Advances in Intelligent Systems and Computing 810

Brijesh Iyer · S. L. Nalbalwar
Nagendra Prasad Pathak *Editors*

Computing, Communication and Signal Processing

Proceedings of ICCASP 2018

Springer

Advances in Intelligent Systems and Computing

Volume 810

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing such as: computational intelligence, soft computing including neural networks, fuzzy systems, evolutionary computing and the fusion of these paradigms, social intelligence, ambient intelligence, computational neuroscience, artificial life, virtual worlds and society, cognitive science and systems, Perception and Vision, DNA and immune based systems, self-organizing and adaptive systems, e-Learning and teaching, human-centered and human-centric computing, recommender systems, intelligent control, robotics and mechatronics including human-machine teaming, knowledge-based paradigms, learning paradigms, machine ethics, intelligent data analysis, knowledge management, intelligent agents, intelligent decision making and support, intelligent network security, trust management, interactive entertainment, Web intelligence and multimedia.

The publications within “Advances in Intelligent Systems and Computing” are primarily proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

More information about this series at http://www.springer.com/series/11156

Brijesh Iyer · S. L. Nalbalwar
Nagendra Prasad Pathak
Editors

Computing, Communication and Signal Processing

Proceedings of ICCASP 2018

Springer

Editors
Brijesh Iyer
Department of Electronics and Telecommunication Engineering
Dr. Babasaheb Ambedkar Technological University
Lonere, Raigad, Maharashtra, India

Nagendra Prasad Pathak
Department of Electronics and Communication Engineering
Indian Institute of Technology Roorkee
Roorkee, Uttarakhand, India

S. L. Nalbalwar
Department of Electronics and Telecommunication Engineering
Dr. Babasaheb Ambedkar Technological University
Lonere, Raigad, Maharashtra, India

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-981-13-1512-1 ISBN 978-981-13-1513-8 (eBook)
https://doi.org/10.1007/978-981-13-1513-8

Library of Congress Control Number: 2018947490

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Dr. Babasaheb Ambedkar Technological University, Lonere-402103, is a State Technological University of Maharashtra State in India. Over the years, the Department of Electronics and Telecommunication Engineering of this university has been organising faculty and staff development programmes and continuing education programmes and workshops. In the year 2013, the department had taken a new initiative to organise international conferences in the areas of signal/image processing, RF and microwave engineering, and IoT. The "ICCASP" series is an outcome of this initiative.

Keynote lectures, invited talks by eminent professors and panel discussion of the delegates with the academicians and industry personnel are the key features of third ICCASP 2018. We have received a great response in terms of the quantity and quality of papers. The conference had adopted a "double-blind review" process to select the papers with a strict plagiarism verification policy. Hence, the selected papers are the true record of research work in their edict.

We are thankful to the reviewers and session chairs and rapporteurs for their support. We also thank the authors and the delegates for their contributions and presence.

We are extremely grateful to Hon Vice-Chancellor, Prof. Vilas G. Gaikar, for his patronage and support from time to time. Financial support for this activity from TEQIP-III is gratefully acknowledged. Finally, we have no words to thank all our colleagues in the department, members of various committees, all the student volunteers, research scholars and alumni without whose unflagging enthusiasm and diligent efforts this conference would not have seen the light of day.

We are pledged to take ICCASP series to greater heights in the years to come with the aim to put forward the need-based research and innovation. To conclude, we would like to express our feelings with the following quote:

असतो मा सद्गमय। तमसो मा ज्योतिर्गमय ।।

(O Lord) Keep me not in the Unreality, but lead me towards the Reality…
(O Lord) Keep me not in the Darkness, but lead me towards the Light.

Thank you one and all.

Lonere, Maharashtra, India	Dr. Brijesh Iyer
Lonere, Maharashtra, India	Dr. S. L. Nalbalwar
Roorkee, India	Dr. Nagendra Prasad Pathak

Contents

About the Editors

Brijesh Iyer holds a Ph.D. in RF and communication engineering from the Indian Institute of Technology Roorkee. Currently, he is Associate Professor in the Department of Electronic and Telecommunication Engineering, Dr. Babasaheb Ambedkar Technological University, Lonere, Raigad. His research interests include nanophotonics/RF front-end design, pervasive healthcare system design and allied signal/image processing. He has published research papers in several peer-reviewed journals and conference proceedings and authored two books. He is a reviewer for many high-impact international journals and actively serves in various committees, e.g. IEEE MTTS, IETE, CSI, IAENG and ISTE.

S. L. Nalbalwar is Professor and Head of the Department of Electronic and Telecommunication, Dr. Babasaheb Ambedkar Technological University, Lonere, Raigad. He received his Ph.D. in signal processing from the Indian Institute of Technology Delhi. His research interests include signal matched filter banks, design, characterisation and process modelling. He has organised many international conferences in his field and is an active member of many professional bodies, e.g. IETE (M123221), CSI (LM54758), ISTE (LM17072), IE (AM0820973), ISCEE (LM212) and IEEE (M80415950). He has published many papers in high-impact journals.

Nagendra Prasad Pathak is Associate Professor at the Indian Institute of Technology Roorkee. He completed his Ph.D. in millimetre-wave integrated circuits at IIT Delhi, after receiving his M.Tech. in electronics engineering from the University of Allahabad. He has developed many courses at IIT Roorkee, e.g. RF and mixed-signal circuits, communication systems and techniques, optical communication and microwave/millimetre-wave integrated circuits. He has worked on many projects as a principal investigator and has published in many journals and conference proceedings. In addition, he has developed two key technologies with applications in defence, health care, disaster management, medicine and veterinary sciences.

Realization of Bandpass Filter Based on Spoof Surface Plasmon Polariton Technique at Microwave Frequency

Gaurav Mittal and Nagendra Prasad Pathak

Abstract Spoof surface plasmon polaritons (SSPPs) are a form of electromagnetic surface wave which, share similar behavior with surface plasma polariton (SPP). The dispersion relation of SSPP is regulated by the geometry of the corrugation using plasmonic metamaterial. In this paper, the SSPP transmission line having double side corrugated strip and bandpass filter which is composed of two opposite oriented single side corrugated strips coupled to one double side corrugated strips are discussed. The re-configurability aspects of SSPP structures are also explored.

Keywords Spoof surface plasmon polariton · Dispersion relation
Plasmonic metamaterial · Bandpass filter

1 Introduction

Recently, plenty of works have demonstrated that the highly confined surface electromagnetic (EM) waves, named spoof SPP or designer SPP, could be supported by plasmonic metamaterial, which consist of a textured metal surface with sub wavelength scaled grooves or dimples [1–5]. The surface plasmon frequency and the SPP like dispersion properties of the spoof SPP could be scaled down to the THz or microwave region by using these plasmonic metamaterial. In this paper, SSPP transmission line based filters using plasmonic metamaterial is discussed. Proposed filter is fed by a transducer composed of a CPW line with a flaring ground.

G. Mittal (✉)
Defence Electronics Applications Laboratory, DRDO, Dehradun 248001, India
e-mail: gauravmittal.drdo@gmail.com

N. P. Pathak
Indian Institute of Technology Roorkee, Roorkee 247667, India
e-mail: nagppfec@iitr.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_1

2 SSPP Transmission Line

In this paper, two types of SSPP transmission line (SSPP-TL) is used which are single side corrugated strip by asymmetric unit cell and double side corrugated strips by symmetric unit cell. Figure 1a, b shows the both symmetric and asymmetric unit cell [6]. The dimension are a = 1 mm, b = 2 mm, c = 5 mm, d = 5 mm. Figure 1c shows the dispersion curve comparison using eigen mode analysis in CST microwave studio. It is observed that the cutoff frequency of the symmetric unit cell is a little bit lower than that of asymmetric unit cell at around 11 GHz.

Recently Tham Yap Fung et al. have presented a hybrid waveguide as shown in Fig. 2 which allows broadband conversion of guided wave to SSPPs and vice versa [7, 8]. The structure shows three sections (i) CPW input (ii) mode converter (iii) SSPP Tx line section. The first part is CPW for the purpose to feed or receive electromagnetic fields. To achieve the 50 Ohm, the CPW parameters are chosen. The second part is a transition section between CPW and the SSPP waveguide, a symmetric periodic corrugated structure. The transition structure consists of flaring ground which is designed to match the impedance and gradient groves are used to match the momentum. The third part is the main periodic SSPP transmission line.

The structure is modeled and simulated in CST software tool with optimized input and output transition. The RT duroide substrate is used for fabrication. The fabricated SSPP transmission line is shown in Fig. 3a. Figure 3b shows the comparison having S-parameter characteristic of simulated and measured results of structure shown in Fig. 2a.

This hybrid mode SSPP transmission line act as a broadband bandpass filter having passband approx (2.1–10.0) GHz which allows signal of a certain band of frequencies to transmit and reject signals of frequencies outside the band.

3 Bandpass Filter Using Coupled SSPP-TL

On the basis of the concept of microwave network, the coupling part can be described by a four port network as shown in the Fig. 4[7]. Bandpass filter is designed using symmetric corrugated strip coupled with two opposite oriented asymmetric corrugated strips [9]. In this way two coupled resonator circuit having four port networks are placed back to back manner. Bandpass filter based on coupling of SSPP-TL is realized.

The design is fabricated on RT Duroid substrate. The fabricated bandpass filter is shown in Fig. 5a, b shows the S-parameter characteristic of simulated and measured results of structure shown in Fig. 5a. The realized bandpass filter has Passband lies between (6.0–7.8) GHz and measured insertion loss is less than 2.0 dB.

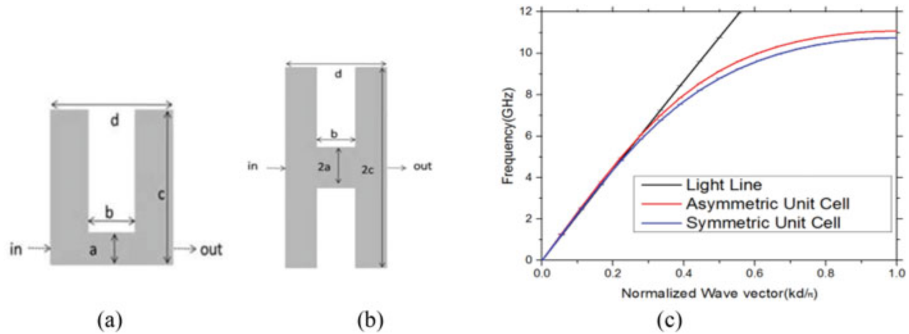


Fig. 1 Unit cell **a** asymmetric **b** symmetric **c** dispersion curve comparison

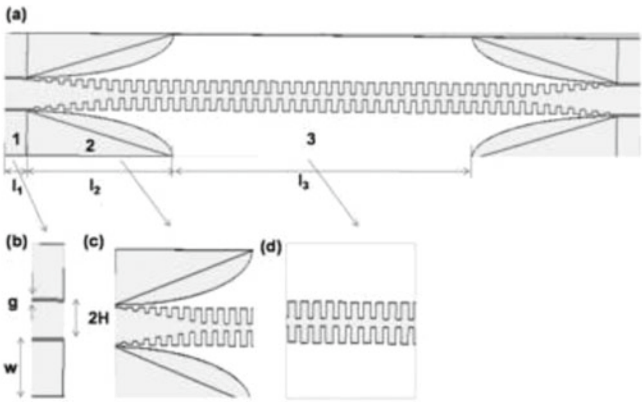


Fig. 2 **a** SSPP Transmission line **b** CPW input **c** mode converter **d** SSPP Tx line

4 Reconfigurable Bandpass Filter

In the previous sections, the SSPP transmission line based filter is discussed. In this section, it will be shown that how the bandwidth or band of the transmission of SSPP structure could be changed. This can be implemented by varying the structure and physical properties of the structure. The filter passband can be change through below mentioned four ways by changing the following:

- Unit cell
- Coupling elements
- Coupled Structure
- Dielectric constant of material

4.1 Changes in Unit Cell

The unit cell as mentioned in Fig. 1b is modified as shown in Fig. 6a having capacitive behavior [10]. The dimension are e = 1 mm, f = 5 mm, g = 3 mm, h = 5 mm. Figure 6b shows the series resonant circuit [11]. Resonant circuits are realized by suitably combining inductor and capacitor. By introducing inductive element in series with unit cell having capacitive behavior, bandpass filter phenomenon can be achieved.

The modified unit cell (Fig. 6a) is further modified using thin microstrip line or inductive element as shown in Fig. 6c. The unit cell (Fig. 6c) is utilized as LC resonator in the filter design. The normalized dispersion curve comparison of Fig. 1b and modified unit cell as Fig. 6a, c and light line are shown in Fig. 7.

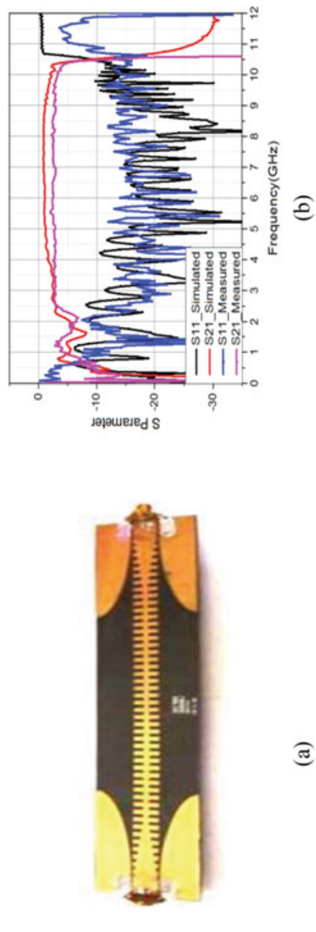


Fig. 3 **a** Fabricated SSPP transmission line **b** simulated and measured S—parameter

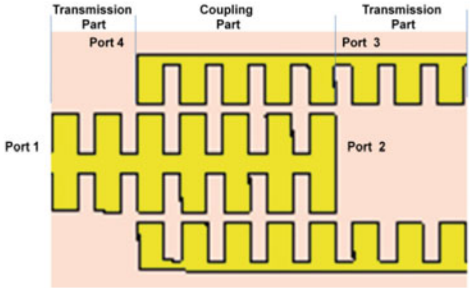


Fig. 4 Coupled resonator phenomenon

As per the dispersion curve shown in Fig. 7, (i) The cutoff frequency for both unit cell mentioned in Figs. 1b and 6a are similar behavior and (ii) The cutoff frequency for modified unit cell having inductive element is much lower than the reported unit cell.

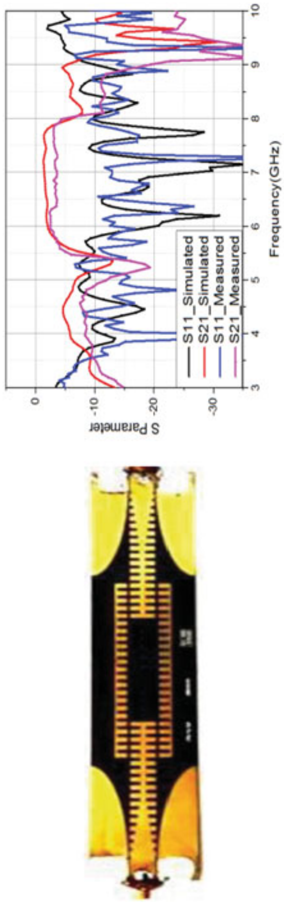
Now, new structure is designed using modified unit cell (Fig. 6c) having order n = 5. The modeling is done in CST as shown in Fig. 8a. Figure 8b shows the simulated S-parameter characteristic having comparison of results between structure of Figs. 3a and 8a. Passband approximately (2.1–6.2) GHz is achieved by structure (Fig. 8a) compare to previous structure shown in Fig. 3a having passband approximately (2.1–10.0) GHz. So pass band at high band side can be decreased or bandwidth can be decreased by new unit cell.

Another way of analysis is that introduction of inductance is actually a defect in the periodic structure. So by introducing the defects in periodic structure, we can achieve modified passband.

4.2 Coupling Elements

On the basis of the frequency selective SSPP structure discussed in Fig. 5a, new structure is designed by increasing the order of coupling element from n = 5 to n = 23 so that SSPP dual-band bandpass filter can be realized [12]. The fabricated SSPP dual-band bandpass filter is shown in Fig. 9a. Figure 9b shows the S-parameter characteristic of simulated and measured results of structure shown in Fig. 9a.

The Band-1 frequency lies between (6.0–6.80) GHz and Band—2 frequency lies between (7.8–8.40) GHz and insertion loss is less than 2.5 dB. So compare to result shown in Fig. 5b having single passband is split into two passband by increasing the number of coupling elements.



(a) (b)

Fig. 5 **a** Fabricated bandpass filter **b** simulated and measured S—parameter results

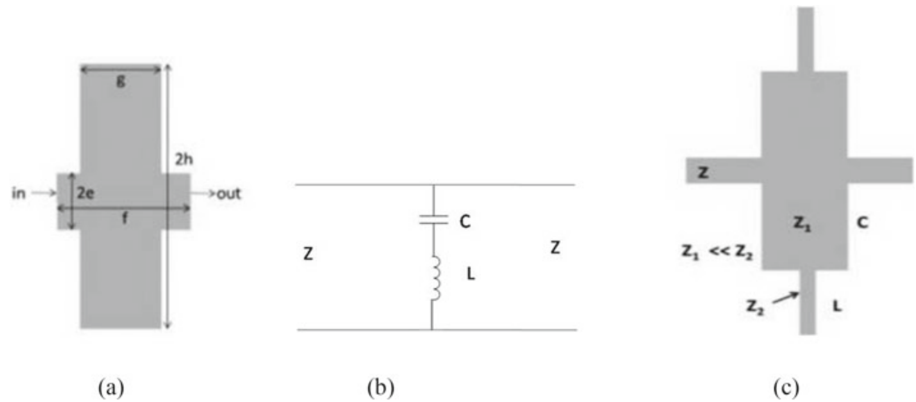


Fig. 6 **a** Proposed symmetric unit cell **b** series LC resonator circuit in lumped domain **c** series resonator circuit in distributed domain

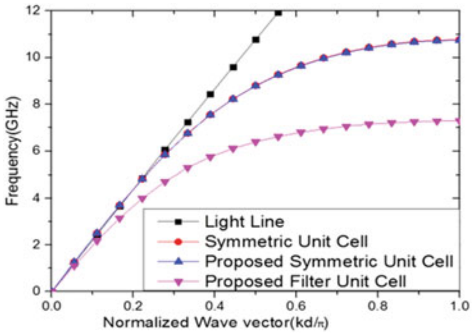


Fig. 7 Dispersion curve comparison of modified unit cell

4.3 Change in Coupled Structure

New structure is designed by keeping symmetric corrugated strip coupled with two opposite oriented symmetric corrugated strips instead of asymmetric. Now SSPP bandpass filter bandwidth become broader compare to design discussed in Fig. 5a. Modeling of structure in shown in Fig. 10a. Figure 10b shows the compared simulated S-parameter characteristic having comparison of results between structure of Fig. 5a and 10a.

Passband approximately (6.0–8.5) GHz is achieved compare to previous passband approximately (6.0–7.8) GHz. So passband at high passband side can be increased or bandwidth is increased.

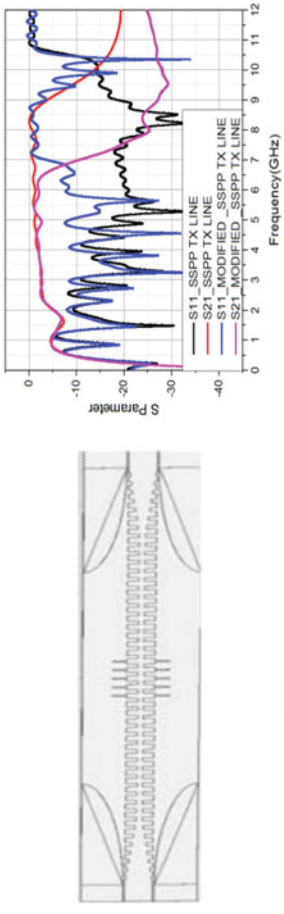
4.4 Dielectric Constant of Material

Passband of bandpass filter can be changed by varying the dielectric constant of substrate for design showing in Fig. 5a. The compared simulated result of substrate having dielectric constant of 2.65 and 2.2 having same thickness are shown in Fig. 11.

Figure 11 Red line ($\varepsilon_r = 2.2$) shows the passband of (6.0–7.8) GHz, BW = 1.8 GHz and pink line ($\varepsilon_r = 2.65$) shows the passband of (5.5–7.0) GHz, BW = 1.5 GHz. So the passband is shifted down and bandwidth decreased for higher dielectric constant having same design.

5 Comparison of Results

In this paper earlier reported work is reproduced/discussed in Sects. 2 and 3. To achieve the reconfigurability, the modification is proposed in Sect. 4. In this section, brief comparison table is shown in Table 1 to summarize the results.



(a) (b)

Fig. 8 **a** SSPP-TL modified waveguide **b** simulated S—parameter results

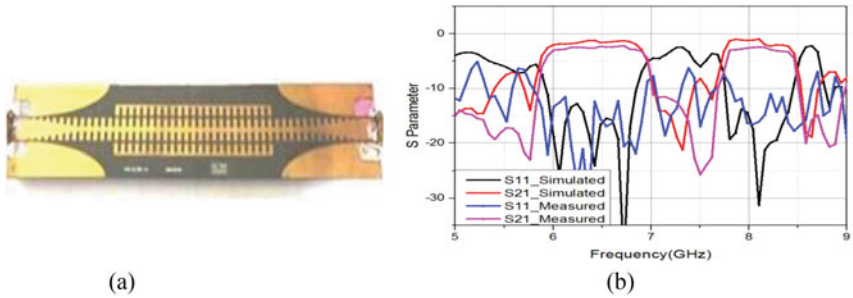


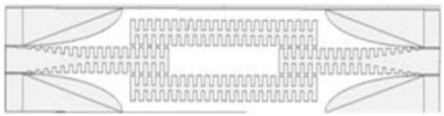
Fig. 9 **a** Dual-band bandpass filter **b** simulated and measured S—parameter results

Table 1 Brief comparison table of results

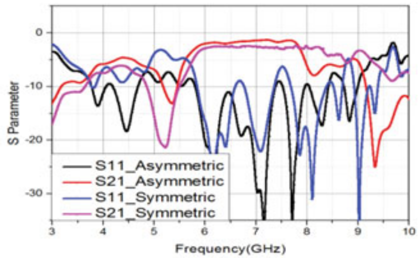
Structure	Variation	Pass band	Remark
Sections 2 and 4.1			
Figure 3a/Sect. 2	Reported/reproduced	(2.1–10) GHz	Proved/reference
Figure 8a/Sect. 4.1	Unit cell	(2.1–6.2) GHz	Bandwidth decreased
Sections 3 and 4.2, 4.3, 4.4			
Figure 5a/Sect. 3	Reported/reproduced	(6.0–7.8) GHz	Proved/reference
Figure 9a/Sect. 4.2	Increased number of coupled unit cell	(6.0–6.8) GHz, (7.8–8.4) GHz	Dual-pass band
Figure 10a/Sect. 4.3	Coupled unit cell, asym. to sym.	(6.0–8.5) GHz	Bandwidth increased and band up-shifted
Figure 5a/Sect. 4.4	Dielectric constant	(5.5–7.0) GHz	Bandwidth decreased and band down-shifted

6 Conclusion

The existence of surface electromagnetic modes in corrugated surfaces of perfect conductors is explored. The main loss in SSPP structure is due to transition having CPW feed. The proposed filter is single layer circuit so may be further used in multi layer structures in the future for higher integration. The demonstration of the manipulation of SSPPs on the hybrid waveguide by varying the structure geometry and physical properties of the hybrid waveguide implies that SSPPs could be highly controllable.



(a)



(b)

Fig. 10 **a** Modeling of modified coupled resonator filter **b** simulated S—parameter

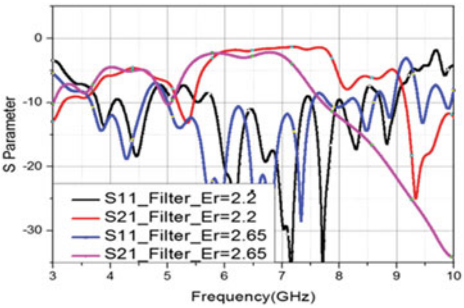


Fig. 11 Simulated S—parameter results

References

1. Shvets, G., Tsukerman, I.: Plasmonics and Plasmonicmetamaterial: analysis and application. World Scientific (2012)
2. Liu, X., Zhu, L., Feng ,Y.: Spoof surface plasmon-based bandpass filter with extremely wide upper stopband physics. Chinese Phys. B **25**, 034101 (2016)
3. Fernandez-Dominguez, A.I., Martin-Moreno, L., Garcia-Vidal, F.J., Andrews, S.R., Maier, S.A.: Spoof surface plasmon polariton modes propagating along periodically corrugated wires. IEEE J. Sel. Top. Q. Electron. **14**, 1515 (2008)
4. Ye, L., Xiao, Y., Liu, Y., Zhang, L., Cai, G., Liu, Q.H.: Strongly confined Spoof surface plasmon polariton waveguiding enabled by planar staggered plasmonic waveguides. Sci. Rep. 38528 (2016)
5. Pendry, J.B., Martin-Moreno, L., Garcia, F.J.: Mimicking surface plasmons with structured surfaces. Science, 847 (2004)
6. Kianinejad, A., Chen, Z.N., Qiu, C.-W.: Design and modeling of spoof surface plasmon modes-based microwave slow-wave transmission line. IEEE Trans. MTT, 63, 1817 (2015)
7. Zhao, L., Zhang, X., Wang, J., Yu, W., Li, J., Su, H., Shen, X.: A novel broadband band-pass filter based on spoof surface plasmon polaritons. Sci. Rep. 36069 (2016)
8. Yin, J.Y., Ren, J., Zhang, H.C., Zhang, Q., Cui, T.J.: Capacitive—coupled series spoof surface plasmon polaritons. Scientific Report,24605(2016)
9. Yin, J.Y., Ren, J., Zhang, H.C., Pan, B.C., Cui, T.J.: Broadband frequency selective spoof surface plasmon polariton on ultrathin metallic structure. Sci. Rep. 08165 (2015)
10. Mittal, G., Pathak, N.P.: Hybrid mode transmission line and bandpass filter implementation using plasmonics metamaterial at microwave frequency. ICIIS, Roorkee, IEEE Xplore Digital Library, 207–210. https://doi.org/10.1109/ICIINFS.2016.826936.2016. 3–4 Dec 2016
11. Bhatt, B., Koul, S.K.: Stripline like Transmission Lines for Microwave Integrated Circuits. New Age International publishers, vol. 601 (2007)
12. Mittal, G., Pathak, N.P.: Techniques of reconfigurable bandpass filter for realization of adaptable RF system and proposed methodology. In: INDICON, Roorkee, 15–17 Dec 2017

Design of Spoof Surface Plasmon Polaritons Based Transmission Line at Terahertz Frequency

Rahul Kumar Jaiswal, Nidhi Pandit and Nagendra Prasad Pathak

Abstract In this paper, we report a plasmonic metamaterial, i.e., spoof surface plasmon polaritons based back to back broadband transition at terahertz frequency. Also we have designed another structure using a unit cell that is made up of by combining three SSPP strip together. This structure shows a way to realize stopband within the operating frequency of spoof surface plasmon polaritons. Using the new type of unit cell disturbs the surface impedance matching and thus gives band stop in the transmission spectrum of SSPP. The first design of transition has reflection coefficient less than −10 dB and transmission loss is less than 5 dB in 0.1–0.8 THz range of frequency. The second designed structure shows strop band from 0.569 to 0.6124 THz and band pass is maintained from 0.1 to 0.569 THz and from 0.6124 to 0.6516 THz. Reflection coefficients in the band-pass region is less than −10 dB and transmission loss is less than 8 dB while in the band stop region reflection coefficient is −3 dB and transmission coefficient is −24 dB has been obtained. Such type structures will show promising application in plasmonic device and systems.

Keywords Terahertz · Spoof surface plasmon polaritons (SSPP) · Plasmonic metamaterial · Dispersion · Band pass · Band stop

---

R. K. Jaiswal · N. Pandit · N. P. Pathak (✉)
Radio Frequency Integrated Circuits and System Laboratory,
Department of Electronics and Communication Engineering, Indian Institute of Technology Roorkee, Roorkee 247667, Uttarakhand, India
e-mail: nagppfec@iitr.ac.in; nagppfec@gmail.com

R. K. Jaiswal
e-mail: rahulj.iitr@gmail.com

N. Pandit
e-mail: npandit7@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_2

1 Introduction

Surface plasmon polaritons (SPP) becomes very good candidate to design plasmonic circuits and device due to its highly localized EM modes at the interface of metal—dielectric. Such type of mode propagates along the interface of two materials only at optical frequency. Conventional optical devices suffer from diffraction limit problem which restrict the feature size of optical circuit and hence no further miniaturization is possible. SPP allows the subwavelength wave propagation and considered as a solid venue for future developments of highly miniaturized optical circuits [1–6]. Thus surface plasmon polaritons have found the application in on-chip integrated optics, super resolution imaging and Nano imaging, subwavelength scale optical circuits, biosensors, photovoltaic, Nano-antenna graphene based devices, etc. [7–13]. However, this kind of behavior is not available at Terahertz or microwave frequencies because metal behaves like highly electrical conductor at these frequencies due to very high value of imaginary part of dielectric constant. Metal behaves like plasma with negative permittivity only beyond of infrared frequency band. Pendry and coworkers have proposed the plasmonic metamaterial which support the modes that shows the similar characteristics as that SPP at optical frequency. Thus realization of the SPP like modes is become possible at THz and Microwave frequency [14, 15]. The characteristics of such type of plasmonic metamaterial are geometry dependent hence it is termed as spoof or designer surface plasmon polaritons. When highly conducting metal surface is corrugated with holes or grooves then that type of periodically corrugated metallic surface supports SPP like modes at these frequencies. The dispersion relation and cut-off frequency of such type of structure can be tuned by changing the geometry parameter of corrugated subwavelength grooves. This type of plasmonic metamaterial can be used to guide and manipulation of the EM waves at subwavelength scale at THz or Microwave frequency. Recently, a no of work related to design of microwave and terahertz circuits and devices has been presented. To convert the guided modes into Spoof SPP mode, various transition designs have been presented in [16–19]. Also design of splitters [20], filters [21–24] and excitation of antenna [25, 26] have been developed.

In this paper we have designed two structures using two type of unit cells; type-1 and type-2. One structure which employ type-1 unit cell shows complete pass band from frequency 0.1–0.8 THz and other structure which employ type-2 unit cell shows a stop band within the operating frequency of spoof SPP. The dispersion characteristics of spoof SPP structure is numerically calculated by considering the finite conductivity of metal and losses associated with dielectric substrate. Broadband transition with gradient groove between microstrip to SSPP are designed to match momentum of the two mode and conversion from QTEM mode of microstrip to spoof SPP mode is obtained.

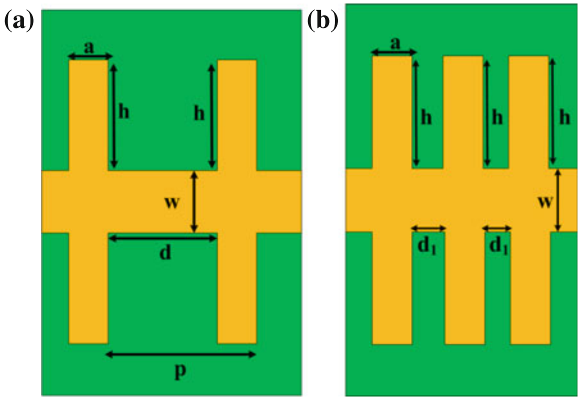


Fig. 1 Schematic of unit cell **a** conventional unit cell (type 1 unit cell), **b** Proposed unit cell (type 2 unit cell)

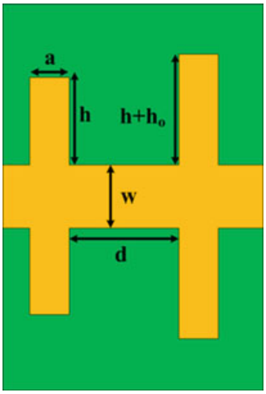


Fig. 2 Matching and conversion unit

2 Dispersion Behavior Analysis of Symmetrically Corrugated Spoof SPP Unit Cell

To analyze the dispersion behavior of spoof SPP at terahertz frequency, first we have designed a unit cell of conventional rectangular shape with gap d as shown in Fig. 1a. Then a new unit cell has been designed by rearranging three SSPP strip together with gap width d_1 as shown in Fig. 1b. Figure 2 shows the conversion unit cell with groove height grows with equal step of $h_{o.}$ The complete section of matching and conversion has been shown in Fig. 3 in which groove height gradually increases from h_1 *to* h_7.

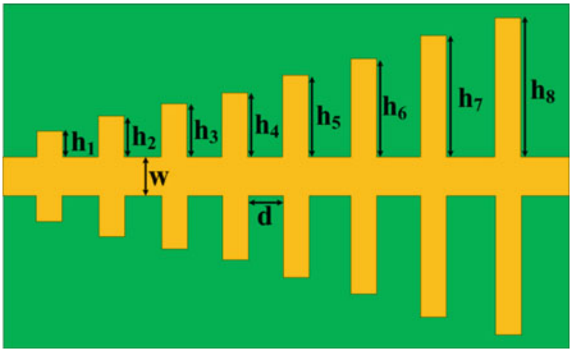


Fig. 3 Schematic of complete conversion section from QTEM mode of microstrip to SSPP mode

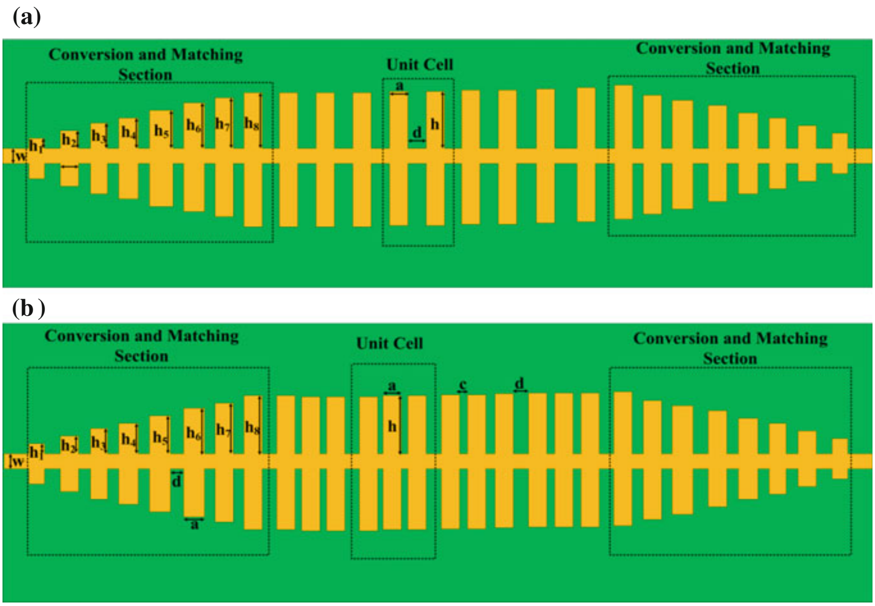


Fig. 4 Schematic of back to back transition from QTEM mode of microstrip to SSPP mode at terahertz frequency for, **a** type 1 unit cell, **b** type 2 unit cell

3 Design of Back to Back QTEM Mode to SSPP Mode Transition

Figure 4 shows the schematic of designed structures. Figure 4a shows the transition (uses the type-1 unit cell) which convert QTEM mode of microstrip to SSPP mode and Fig. 4b show the structure which is designed using type-2 unit cell.

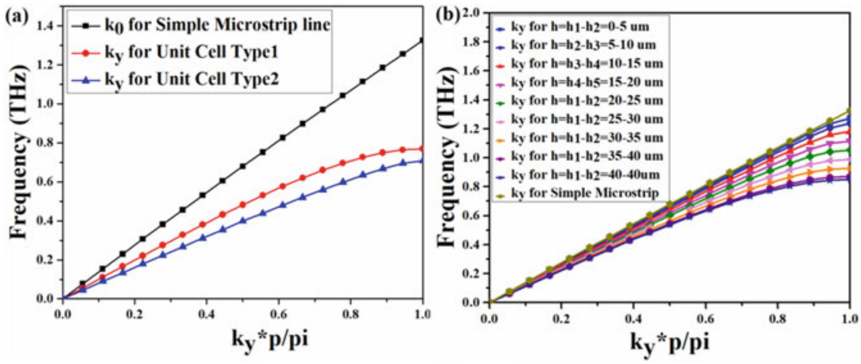


Fig. 5 **a** Dispersion characteristics for unit cell as shown in Fig. 1a, b, **b** dispersion characteristics that shows the momentum matching and conversion

4 Results and Discussion

Figure 5a shows the dispersion relation for designed unit cells. The numerical simulation has been performed using CST Microwave Studio. In Fig. 5a red color curve shows dispersion relation for type-1 unit cell and blue color line shows the dispersion for type-2 unit cell. Black color line shows the propagation wave vector for freely propagating wave that is propagation wave vector of simple microstrip line without any grooves. The wave vectors *ky* of SSPP unit cell (red and blue curve) are very much larger than the wave vector of freely propagating wave vector *ko* (black line) as depicted from Fig. 5a, hence there is a mismatch between wave vectors *ky* and *ko*. Hence to design a complete conversion from QTEM mode of microstrip to SSPP, we need to match the momentum (matching between wave vectors *ky* and *ko*) and this matching has been performed by using gradient corrugated grooves as shown in Fig. 3.

As the height of grooves increases from h_1 to $h_{7,}$ the wave vector also increases and it deviates from the wave vector of freely propagating wave vector and finally it matches to wave vector of SSPP unit cell as shown in Fig. 5b. Then a convertor from QTEM mode of microstrip to SSPP mode has been designed as shown in Fig. 4. Figure 6 shows the reflection and transmission coefficient for designed structures (Fig. 4a, b). Figure 6a shows the S-parameter for designed transition using type-1 unit cell, which shows complete pass band from 0.1 to 0.8 THz with reflection and transmission coefficient are less than −10 and −5 dB. Figure 6b shows the S-parameter for designed structure using type-2 unit cell, which shows stop band from 0.569 to 0.6124 THz with reflection and transmission coefficient are around −3 dB and −24 dB respectively and pass band from 0.1 to 0.569 THz and 0.6124–0.6516 THz with reflection and transmission coefficient less than −10 dB and −8 dB respectively.

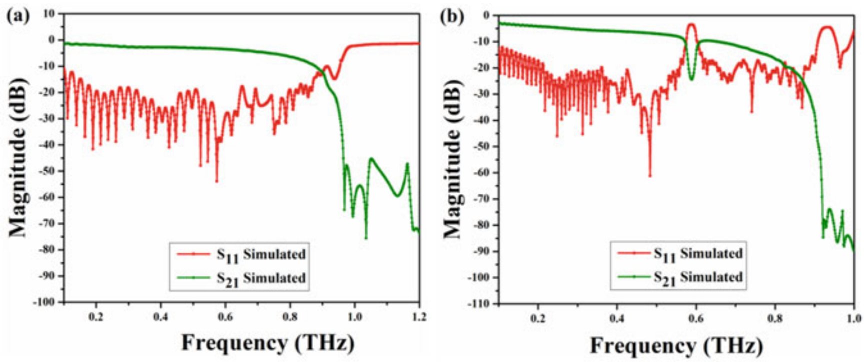


Fig. 6 Simulated magnitude of reflection and transmission coefficient for, **a** designed back to back transition using unit cell as in Fig. 1a, b using unit cell as in Fig. 1b

5 Conclusion

In this paper, the design and analysis of transition from QTEM mode of microstrip line to SSPP mode using type-1 unit cell has been discussed at terahertz frequency. When we design structure using unit cell of type-2 then we get stop band within the operating frequency range of SSPP due to change in the surface impedance. Simulated results of dispersion behavior of unit cells and reflection and transmission coefficients for designed transitions using both type of unit cell have been discussed. When we have used type 2 unit cell then we obtain a bandstop region in the transmission spectrum of spoof SPP. Thus the transition geometry presented in this paper can be used to develop other plasmonic circuits using the concept of spoof or designer surface plasmon polariton at terahertz frequency.

References

1. Zayats, A.V., Smolyaninov, I.I., Maradudin, A.A.: Nano-optics of surface palsmon polaritons. Phys. Rep. **408**, 131–314 (2005)
2. Maier, S.A.: Plasmonics: Fundamentals and Applications. Springer Verlag, USA, New York, NY (2007)
3. Atwater, H.A.: The promise of plasmonics. Sci. Am. **296**, 56–62 (2007)
4. Web, S.I., This, S., Press, H., York, N., Nw, A.: Plasmonics: merging photonics and electronics at nanoscale dimensions. Sci. Rev. **311**, 189–194 (2006)
5. Pitarke, J.M., Silkin, V.M., Chulkov, E.V., Echenique, P.M.: Theory of surface plasmons and surface-plasmon polaritons. Rep. Prog. Phys. **1** (2006)
6. Rusina, A., Durach, M., Stockman, M.I.: Theory of spoof plasmons in real metals. Appl. Phys. A Mater. Sci. Process. **100**, 375–378 (2010)
7. Liu, H., Wang, B., Ke, L., Deng, J., Chum, C.C., Teo, S.L., Shen, L., Maier, S.A., Teng, J.: High aspect subdiffraction-limit photolithography via a silver superlens. Nano Lett. **12**, 1549–1554 (2012)

8. Wang, B., Zhang, X., Yuan, X., Teng, J.: Optical coupling of surface plasmons between graphene sheets. Appl. Phys. Lett. **100** (2012)
9. Atwater, H., Polman, A., Kosten, E., Callahan, D., Spinelli, P., Eisler, C., Escarra, M., Warmann, E., Flowers, C.: Nanophotonic design principles for ultrahigh efficiency photovoltaics. AIP Conf. Proc. **1519**, 17–21 (2013)
10. Schmidt, M.A., Lei, D.Y., Wondraczek, L., Nazabal, V., Maier, S.A.: Hybrid nanoparticle-microcavity-based plasmonic nanosensors with improved detection resolution and extended remote-sensing ability. Nat. Commun. **3** (2012)
11. Zhang, S., Xiong, Y., Bartal, G., Yin, X., Zhang, X.: Magnetized plasma for reconfigurable subdiffraction imaging. Phys. Rev. Lett. **106**, 1–4 (2011)
12. Jones, A.C., Olmon, R.L., Skrabalak, S.E., Wiley, B.J., Xia, Y.N., Raschke, M.B.: Mid-IR polasmonics: Near-field imaging of coherent plasmon modes of silver nanowires. Nano Lett. **9** (2009)
13. Anker, J.N., Hall, W.P., Lyandres, O., Shah, N.C., Zhao, J., Van Duyne, R.P.: Biosensing with plasmonic nanosensors. Nat. Material. **7**, 442–453 (2008)
14. Pendry, J.B., Martin-Moreno, L., Garcia-Vidal, F.J.: Mimicking surface plasmons with structured surfaces. Science **305**, 847–848 (2004)
15. Hibbins, A.P., Evans, B.R., Sambles, J.R.: Experimental verification of designer surface plasmons. Sci. Mag. **308** (2005)
16. Zhang, W., Zhu, G., Sun, L., Lin, F.: Trapping of surface plasmon wave through gradient corrugated strip with underlayer ground and manipulating its propagation: J. App. Phys. **106** (2015)
17. Ma, H.F., Shen, X., Cheng, Q., Jiang, W.X., Cui, T.J.: Broadband and high-efficiency conversion from guided waves to spoof surface plasmon polaritons. Laser Photonics Rev. **8**, 146–151 (2014)
18. Kianinejad, A., Chen, Z.N., Qiu, C.-W.: Design and modeling of spoof surface plasmon modes-based microwave slow-wave transmission line. IEEE Trans. Microwave Theory Tech. **63**, 1817–1825 (2015)
19. Liu, X., Zhu, L., Wu, Q., Feng, Y.: Highly-Confined and Low-Loss Spoof Surface Plasmon Polaritons Structure with Periodic Loading of Trapezoidal Grooves, vol. 5. AIP Advances. (2015)
20. Gao, X., Shi, J.H., Shen, X., Ma, H.F., Jiang, W.X., Li, L., Cui, T.J.: Ultrathin dual-band surface plasmonic polariton waveguide and frequency splitter in microwave frequencies. Appl. Phys. Lett. 102, 9–13 (2013)
21. Jaiswal, R.K., Pathak, N.P.: Spoof surface plasmons polaritons based multi-band band pass filter. IEEE APMC Conference (2016) (in press)
22. Jaiswal, R.K., Pathak, N.P.: Development and design of multi-band bandpass filter based on the concept of spoof surface plasmon polaritons. In: IEEE ICIIS Conference (2016) (in press)
23. Jaiswal, R.K., Pandit, N,. Pathak, N.P.: A broadband transition device and multiband band-pass filter using ring resonator based on spoof surface plasmon polaritons at microwavefrequency. IEEE IMarc (2017) (in press)
24. Jaiswal, R.K., Pandit, N., Pathak, N.P.: Design of Multiple Band-Notch Ring Resonator Filter Based on Plasmonic Metamaterial at Microwave Frequency. IEEE iAim, Nov (2017) (in press)
25. Jaiswal, R.K., Pandit, N., Pathak, N.P.: Design, Analysis and Characterization of Designer Surface Plasmon Polaritons—Based Dual Band Antenna. Springer Plasmonics (2017). https://doi.org/10.1007/s11468-017-0622-1
26. Liu, N., Wen, F., Zhao, Y., Wang, Y., Nordlander, P., Halas, N.J., Alu, A.: Individual nanoantennas loaded with three-dimensional optical nanocircuits. Nano Lett. **13**, 142–147 (2013)

Multiband Multimode Filter for Wireless Applications

Nidhi Pandit, Rahul Kumar Jaiswal and Nagendra Prasad Pathak

Abstract This paper reports, design, analysis, and characterization of multimode resonator based multiband band pass filter. In support to the actual response, principle of resonating structure is explained with related mathematical explanations. For validating the concept, a quad-band BPF have been implemented on Neltec substrate and characterized through Keysight vector network analyzer *N9918A*. All the measured and simulated results are in good agreement with each other.

Keywords Multimode · Resonator · BPF · Microstrip

1 Introduction

Filter plays an important role in Microwave receiver system. Resonators are the key element of filters hence translates the performance of BPF. Researchers have proposed a number of methods for improvising filtering response and miniaturizing structure size. Like in [1, 2], SIR concept with uniform λ/4 resonator is used to provide features like multiband response and wide stop-band performance along with miniaturization. Further for miniaturization, some novel multimode resonating structures were proposed in past years [3–8]. Multimode resonators are attractive area of research because it behaves like multi-toned resonator with N number of resonating frequency; hence for required performance reduced the number of resonator by 1/N. Multimode resonator based on stub loading was firstly reported in [3], because of

N. Pandit · R. K. Jaiswal · N. P. Pathak (✉)
Radio Frequency Integrated Circuit Research Laboratory, Department of Electronics and Communication Engineering, Indian Institute of Technology Roorkee, Roorkee 247667, India
e-mail: nagppfec@iitr.ac.in; nagppfec@gmail.com

N. Pandit
e-mail: npandit7@gmail.com

R. K. Jaiswal
e-mail: rahulj.iitr@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_3

its several advantages multimode stub loaded resonator are conveniently used in RF designs [4–8].

In this paper, multimode resonator based multiband microwave filter has been presented. In Sect. 2, principle of resonating structure has explained with mathematical analysis. Section 3, address the experimental and simulation results. Finally conclusion has been drawn in Sect. 4.

2 Proposed Multimode Structure

The proposed multimode resonator configuration is shown in Fig. 1a. Because of symmetry in the proposed resonating structure i.e. Fig. 1a, its behavioral characteristics can be conveniently explained through even-odd-mode analysis.

The even and odd-mode circuits of Fig. 1a are shown in Fig. 1b and Fig. 1c with input admittance of Y_{ine}, Y_{ino} respectively. Y_1, Y_2 and L_{s1}, L_{s2} are admittances and physical length of basic λ/2 stepped impedance multimode resonator. Y_3, Y_4, Y_5, Y_6 correspond to characteristic admittances and L_{s3}, L_{s4}, L_{s5}, L_{s6} of stubs and s and s_1 refer to physical spacing between stubs. For ease of analysis let us assume $s = s_1$ so $\theta_s = \theta_{s1}$ is defined as βs and θ_n is defined as βL_{sn} for $n = 1$ *to* 6.

The resonance condition for the proposed resonator can be achieved and expressed as

$$\begin{gathered}
\mathrm{Y_{ino}} = 0 \\
\text{where} \quad \mathrm{Y_{ino}} = \mathrm{Y_{in}} \rightarrow \mathrm{Y_{in}} = \mathrm{Y_1}\frac{\mathrm{P_1 + jY_1 \tan(\theta_1)}}{\mathrm{Y_1 + jP_1 \tan(\theta_1)}} \\
\text{when} \quad \mathrm{P_1} = -\mathrm{jY_2}\cot(\theta_s) - \mathrm{jY_3}\cot(\theta_3) + \mathrm{jY_6}\tan(\theta_6) \quad \text{for odd mode}
\end{gathered} \tag{1}$$

$$\begin{gathered}
\mathrm{Y_{ine}} = 0 \\
\text{where} \quad \mathrm{Y_{ine}} = \mathrm{Y_{in}} \rightarrow \mathrm{Y_{in}} = \mathrm{Y_1}\frac{\mathrm{P_1 + jY_1 \tan(\theta_1)}}{\mathrm{Y_1 + jP_1 \tan(\theta_1)}} \\
\text{when} \quad \mathrm{P_1} = -\mathrm{jY_2}\cot(\theta_s) - \mathrm{jY_3}\cot(\theta_3) + \mathrm{jY_6}\tan(\theta_6) + \mathrm{T_2} \\
\text{with} \quad T_2 = \mathrm{Y_2}(\mathrm{T_1} + \mathrm{jY_2}\tan(\theta_s))/(\mathrm{Y_2} + \mathrm{jT_1}\tan(\theta_s)), \\
\mathrm{T_1} = (\mathrm{jY_5}/2)\tan(\theta_5) - (\mathrm{jY_4}/2)\cot(\theta_4) \quad \text{for even mode}
\end{gathered} \tag{2}$$

It is clear from (1) to (2) and Fig. 1, that L_1, L_2, L_{s1}, L_s, L_{s3}, L_{s6} affects both the even and odd-mode frequencies while L_{s4}, L_{s5} affects the even-mode only.

3 Simulation and Measured Results with Discussion

Figure 2 shows the layout of proposed BPF that has been obtained by modifying the structural parameters of proposed multimode resonator according to Eqs. (1)–(2).

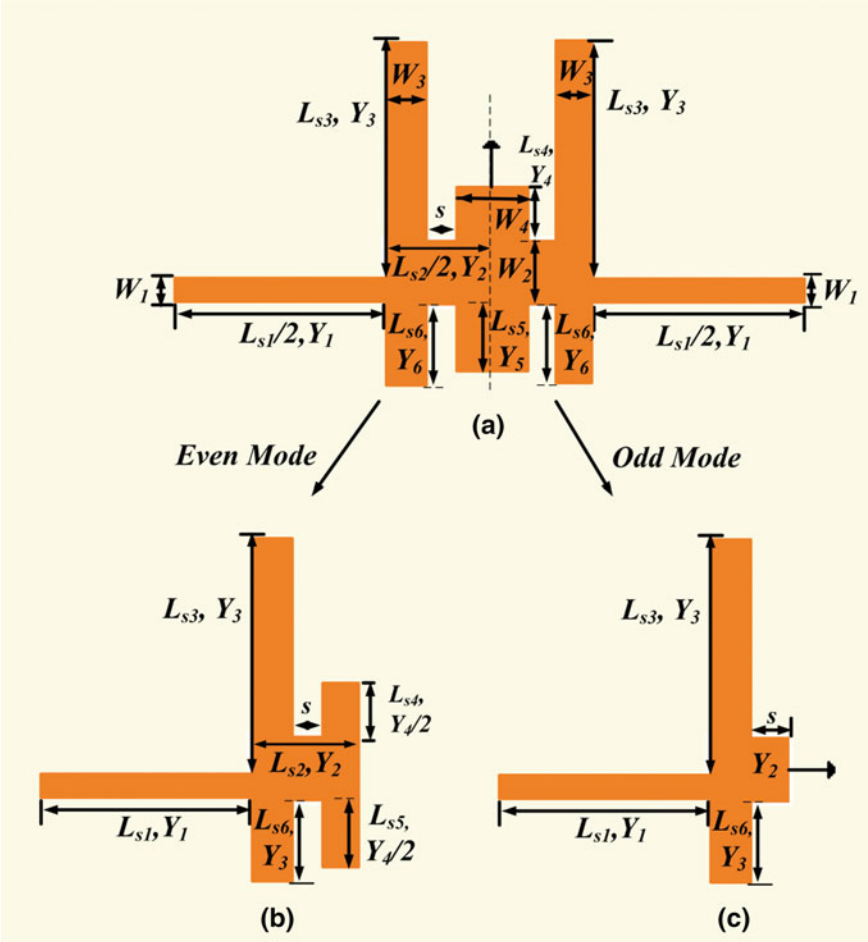


Fig. 1 **a** Proposed multimode resonator and its equivalent. **b** Even-mode and **c** odd-mode circuit

Design Parameters (all in mm): $L_{s1} = 34.62$, $L_{s2} = 18.56$, $L_{s3} = 5$, $L_{s4} = 11.8$, $L_{s5} = 1.4$, $L_{s6} = 4.6$, $L_7 = 10$, $L_8 = 16.5$, $s = 0.25$, $s_1 = 4.5$, $s_2 = 0.2$, $W_1 = 4.85$, $W_2 = 6.75$, $W_3 = 6.9$, $W_4 = 1.3$, $W_5 = 1.36$, $W_6 = 2$, $W_7 = 1$.

For validating the design, prototype of proposed filter structure has fabricated as shown in Fig. 3 with optimized physical parameters on 1.52 mm thick Neltec substrate with dielectric constant 3.38 and loss tangent of 0.0016. Additionally, interdigital feeding line is utilized to achieve better coupling in passband.

The S-parameters of fabricated prototype i.e. S_{11} and S_{21} were measured using Keysight Field-Fox Microwave Analyzer *N9918A* and shown in Figs. 4 and 5. Designed filter have quad-band response with center frequencies 1.74 GHz, 2.64 GHz, 4.1 GHz, 6.8 GHz with measured −3 dB impedance bandwidth of 40 MHz,

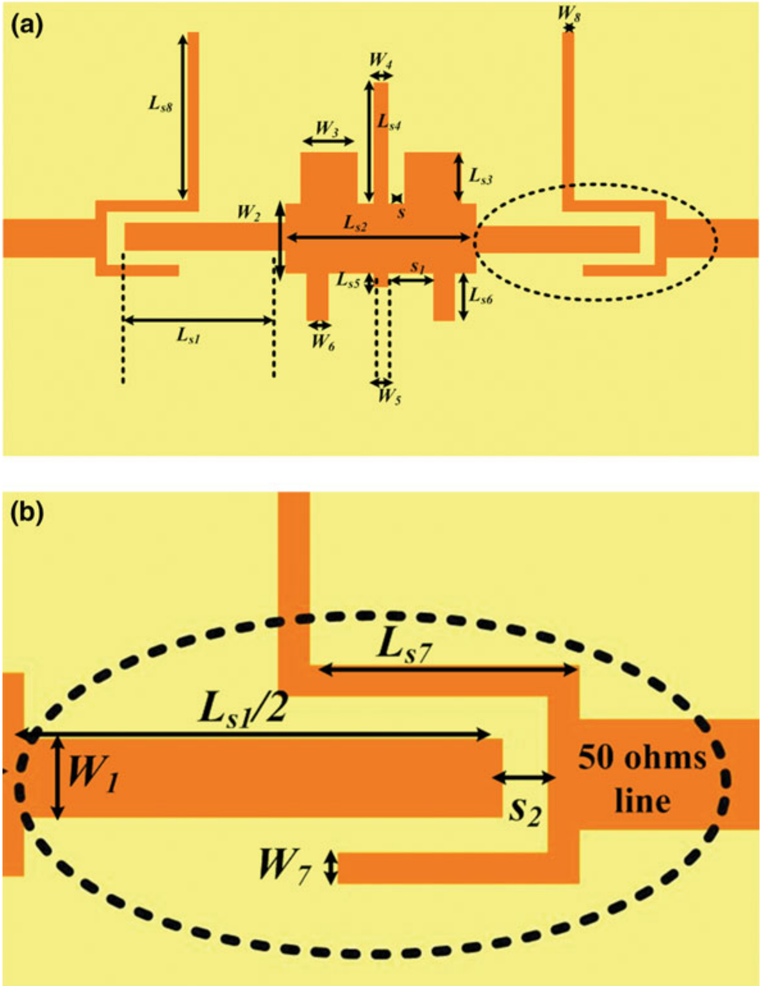


Fig. 2 **a** Layout of proposed quad-band multimode BPF, **b** interdigital feed line used in (**a**)

40 MHz, 220 MHz, 420 MHz, respectively. For the designed filter insertion loss of 2 dB (minimum) and reflection losses better than 10 dB have been observed within the specified frequency bands respectively. Measured results are consistent with simulation results but a slight magnitude difference is mainly caused due to fabrication tolerance. As compared to previously reported works, the proposed filter is provide flexible control on mode coupling hence multiband frequency response with desired characteristics have been obtained.

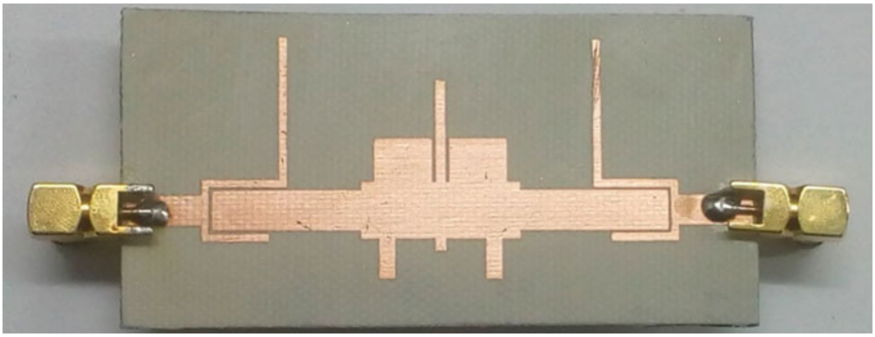


Fig. 3 Fabricated prototype of proposed quad-band BPF

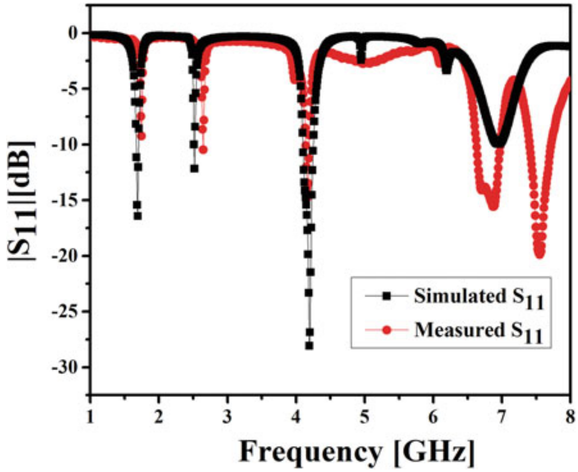
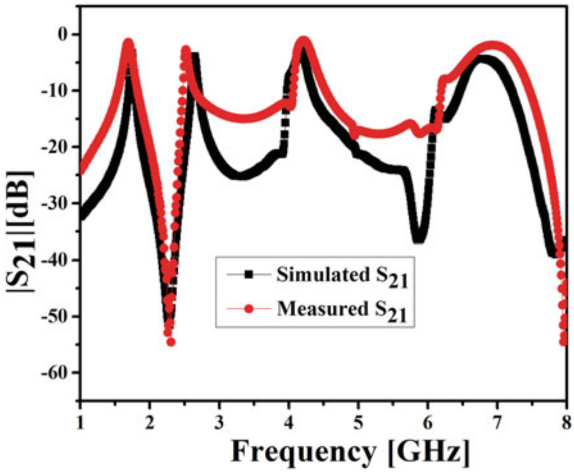


Fig. 4 S_{11}-parameters of fabricated quad-band BPF

4 Conclusion

Design of multiband multimode BPF has been discussed in this paper. Based on the coupling behavior of resonating frequencies different physical parameters have been optimized to get required filter characteristics. Further to achieve tight coupling with in the passband of filter structure, interdigital coupled feed line is used. Verification of design concept has been done through implementation and characterization of hardware prototype for the proposed structures. With feature such as flexible control on resonating modes and low insertion loss; the proposed filter is attractive choice for modern wireless communication system application.

Fig. 5 S_{21} parameters of fabricated quad-band BPF



References

1. Lin, S.-C., Deng, P.-H., Lin, Y.-S., Wang, C.-H., Chen, C.H.: Wide-stopband microstrip bandpass filters using dissimilar quarter-wavelength stepped-impedance resonators. IEEE Trans. Microw. Theory Tech. **549**, 1011–1018 (2013)
2. Zhang, S., Zhu, L.: Synthesis design of dual-band bandpass filters with stepped-impedance resonators. IEEE Trans. Microw. Theory Tech. **61**, 1812–1819 (2013)
3. Lee, J.R., Cho, J.H., Yun, S.W.: New compact bandpass filter using microstrip resonators with open stub inverter. IEEE Microw. Guided Wave Lett. **10**, 526–527 (2000)
4. Li, Y.C., Wong, H., Xue, Q.: Dual-mode dual-band bandpass filter based on a stub- loaded patch resonator. IEEE Microw. Wirel. Compon. Lett. **21**, 525–527 (2011)
5. Zhang, X.Y., Chen, J.-X., Xue, Q., Li, S.-M.: Dual-band bandpass filters using stub-loaded resonators. IEEE Microw. Wirel. Compon. Lett. **17**, 583–585 (2007)
6. He, W., Ma, Z., Chen, C.-P., Anada, T., Kobayashi, Y.: A novel dual-band bandpass filter using microstrip stub-loaded two-mode resonators with source and load coupling. In: IEEE Asia-Pacific Microwave Conference (2008)
7. Gao, L., Zhang, X.Y., Hu, B.-J., Xue, Q.: Novel multi-stub loaded resonators and their applications to various bandpass filters. IEEE Trans. Microw. Theory Tech. **62**, 1162–1172 (2014)
8. Ma, Z., He, W., Chen, C.-P., Anada, T.: Miniaturized high-performance UWB bandpass filter consisting of only one microstrip three-mode resonator. In: IEEE European Microwave Conference (2009)

Design of Graphene-Based THz Antennas

Arun Kumar Varshney, Nagendra Prasad Pathak and Debabrata Sircar

Abstract This paper first reports the design of the tunable graphene-based patch antenna at THz frequencies. After that a tunable graphene-based U slot loaded patch antenna has been designed to increase the bandwidth at THz frequencies. The simulated results for both the antennas are presented for different values of chemical potential.

1 Introduction

Terahertz (THz) frequency band ranging from 0.1 to 10 THz, was one of the least explored regions because lack of materials responding to these frequencies. But in the recent year, Graphene has produced great interest in various academic and industrial researches in THz region. Due to its amazing mechanical, optical, electronics and chemical properties [1–3], graphene is much promising for designing of Nano electronics and Nano photonics devices.

Graphene is a 2D sheet of carbon atoms arranged in a honeycomb structure, which has concerned remarkable interest in electrical and mechanical fields. Graphene is a zero gap semiconductor and the complex conductivity nature of graphene allows the propagation of surface plasmonic modes at optical frequencies. Importantly, this conductivity can be efficiently controlled via a perpendicular bias electric field. The change of surface conductivity will affect the resonance frequency of the reflector. As a result graphene can be used for a variety of applications at THz and optical frequencies, including the possibility of dynamic tuning via the electric field effect.

A. K. Varshney (✉) · N. P. Pathak · D. Sircar
Indian Institute of Engineering and Technology Roorkee, Roorkee 247667, India
e-mail: aruniitr2014@gmail.com

N. P. Pathak
e-mail: nagppfec@iitr.ac.in

D. Sircar
e-mail: debsrfbt@iitr.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_4

In recent time, we reported some graphene-based devices such as amplifiers [4], diodes [5], field effect transistors [6], electro-optical switches [7], mixers [8], filters [9], terahertz antennas [10, 11], absorbers [12], plasmonic Bragg reflectors [13], multimode interferometer [14] and so on. The graphene is also used as an electrode in a hybrid photovoltaic cells [15]. Because the conductivity and losses of graphene can be tuned, the broadband absorber have been proposed by varying the gate voltage [12]. A graphene polarizer has also been proposed in the near-infrared band, which can work as a (TM)/(TE) polarizer when the upper-cladding layer is/not dropped on the waveguide core [16]. A tunable optical delay line has also been proposed because of unique properties of graphene [17].

2 Electronic Model of Graphene

Due to the mono-atomic thickness, graphene can be modeled by a surface conductivity σ. In the absence of applied magnetic fields, the value of conductivity is computed by Kubo's formula [18]. Its expresses that conductivity is depend on temperature T, chemical potential μ_c, frequency ω and scattering rate Γ (Fig. 1).

$$\sigma(\omega, \mu_c, \Gamma, T) = \frac{je^2(\omega - j2\Gamma)}{\pi h^2} \left[\frac{1}{(\omega - j2\Gamma)^2} \int_0^\infty \varepsilon \left(\frac{\partial f_d(\varepsilon)}{\partial \varepsilon} - \frac{\partial f_d(-\varepsilon)}{\partial \varepsilon} \right) d\varepsilon - \int_0^\infty \frac{f_d(-\varepsilon) - f_d(\varepsilon)}{(\omega - j2\Gamma)^2 - 4\left(\frac{\varepsilon}{h}\right)^2} d\varepsilon \right] \quad (1)$$

where $f_d(\varepsilon) = \left(e^{\frac{\varepsilon - \mu_c}{k_B T}} + 1\right)^{-1}$ is the Fermi-Dirac distribution function, $\hbar = h/2\pi$ where h is the Planck's constant, e is an electron charge, and k_B is the Boltzmann's constant. Throughout this work, we use following parameters values T = 1 ps and T = 300 K. The first term in this equation is because of intra band contributions, and the second term is because of inter band contributions. At low frequencies, due to low energy photones, the conductivity of graphene is mainly dominant by the intra band transitions, while at the higher frequencies, the effect of inter band transitions becomes dominant. For the cases considered here the intra band term dominates and can be evaluated as

$$\sigma_{\text{intra}}(\omega, \mu_c, \Gamma, T) = -j \frac{e^2 k_B T}{\pi \hbar^2 (\omega - j2\Gamma)} \left(\frac{\mu_c}{k_B T} + 2 \ln\left(e^{-\frac{\mu_c}{k_B T}} + 1 \right) \right) \quad (2)$$

And the conductivity due to inter band transitions can be evaluated as

$$\sigma_{inter}(\omega, \mu_c, \Gamma, T) = -j \frac{e^2}{4\pi \hbar} \ln\left(\frac{2|\mu_c| - (\omega - j2\Gamma)\hbar}{2|\mu_c| + (\omega - j2\Gamma)\hbar} \right) \quad (3)$$

The graphene optical conductivity can be changed by controlling the Fermi level, i.e., carrier concentration. Thus, the conductivity can be efficiently controlled via a perpendicular bias electric field. As a result, graphene is envisioned for a variety of applications at THz and optical frequencies.

The inductive nature of conductivity allows an infinite sheet of graphene to support TM surface waves also called as Surface Plasmon Polaritons (SPPs). When an infinite sheet of graphene inserted between two dielectric materials, the dispersion equation of the mode is given by

$$\frac{\in_{r1}}{\sqrt{k^2 - \in_{r1} k_0^2}} + \frac{\in_{r2}}{\sqrt{k^2 - \in_{r2} k_0^2}} = \frac{-j\sigma}{\omega \in_0} \tag{4}$$

where $\in_{r1}$ and $\in_{r2}$ are the relative dielectric constant, k_0 is the free space wave number, $k = \beta - j\alpha$ is the guided propagation constant, α is the attenuation constant and β is the guided wave number.

3 Geometry of the THz Antennas Based on Graphene

In many academic and industrial researches, graphene antennas are one of the best techniques for generation and detection of THz radiation. By adjusting the graphene conductivity, a reconfigurable terahertz graphene antenna can allow both frequency tuning and beam steering [10]. In this work, we presented a graphene-based simple patch antenna and U slot loaded patch antenna at THz frequencies. In this work, we used CST microwave studio (electromagnetic simulation tool) to obtain optimized characteristics of the antenna. For designing the antenna, we used a SiO_2 substrate, which have a dielectric constant $\in_r = 3.75$ and very low losses at THz frequencies.

3.1 THz Patch Antenna Based on Graphene

Around a one-half wavelength long section of transmission line is used to design a microstrip patch antenna. A simple patch antenna has been designed at THz frequencies. The geometry of this antenna is optimized using CST microwave studio and shown in Fig. 2a. The patch of the antenna is composed by graphene layer with dimensions $W = 48.66$ µm and $L = 38.54$ µm and its feed by a metallic transmission line of width 0.03 µm (Fig. 2a).

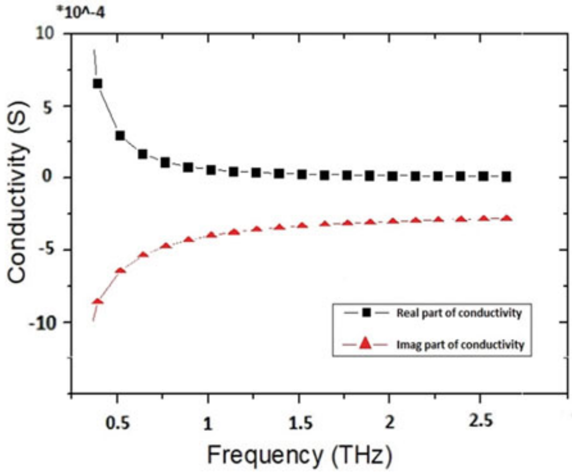


Fig. 1 Graphene complex conductivity for $\mu_c = 0.25\,\mathrm{eV}$, $T = 300\,\mathrm{K}$ and scattering rate $= 1\,\mathrm{ps}$

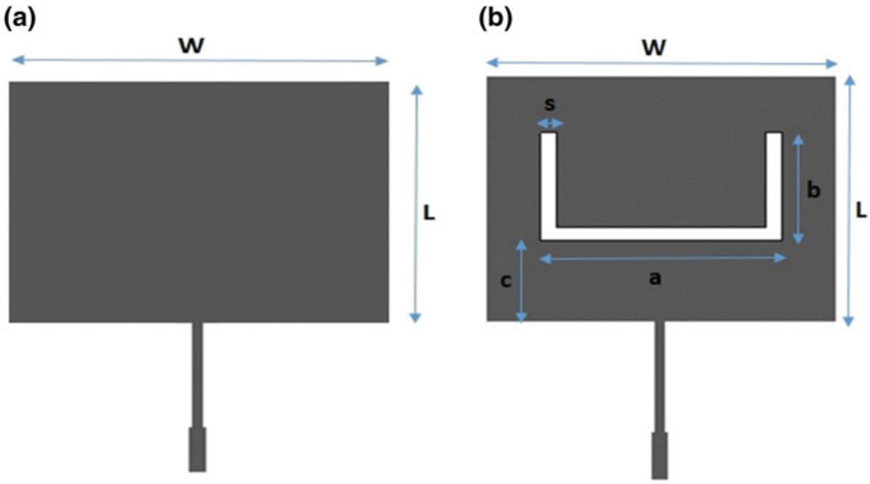


Fig. 2 Geometry of the **a** simple patch antenna **b** U slot loaded patch antenna

3.2 THz U Slot Loaded Patch Antenna Based on Graphene

Several methods have been invented for bandwidth increment of the microstrip patch antennas in addition to other usual methods of decreasing substrate permittivity and increasing patch height. In this section, a simple design for the rectangular U slot Microstrip patch antenna has been presented. The geometry of this antenna is

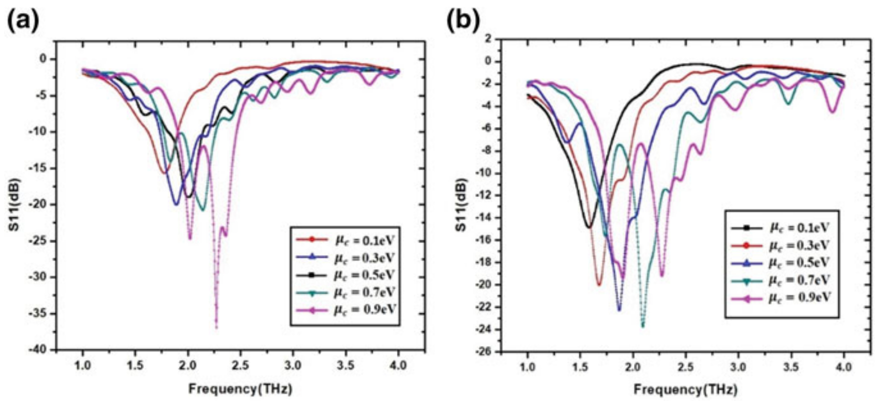


Fig. 3 Simulated return loss for **a** simple patch antenna **b** U slot loaded patch antenna for different values of $\boldsymbol{\mu_c}$

optimized using CST microwave studio and shown in Fig. 2b. The width and length of the patch are 48.66 µm and 38.54 µm respectively and the other design specifications are such as horizontal slot length a = 32 µm, vertical slot length b = 20 µm, parameter c = 13 µm and the width of each slot s = 2 µm. The antenna structure is feed with metallic transmission line of width 0.03 µm as shown in Fig. 2b.

4 Simulation Results for Graphene-Based Antennas

The simulated result for simple patch antenna and U slot loaded antenna, computed for different values of μ_c is shown in Figs. 3 and 4. It is founded a large tuning range with resonant frequencies in the THz frequency band. Figure 5a and Fig. 5b shown the radiation patterns for the simple patch antenna and U slot loaded antenna respectively for μ_c = 0.5 eV.

5 Conclusion

It has been shown that due to its amazing electro-optical property, graphene is a much favorable material for designing of THz antennas. And it has been also shown that its conductivity can be efficiently controlled via a perpendicular bias electric field. It is worth mentioning that graphene can be used for a variety of applications at THz and optical frequencies , including the possibility of dynamic tuning via the electric field

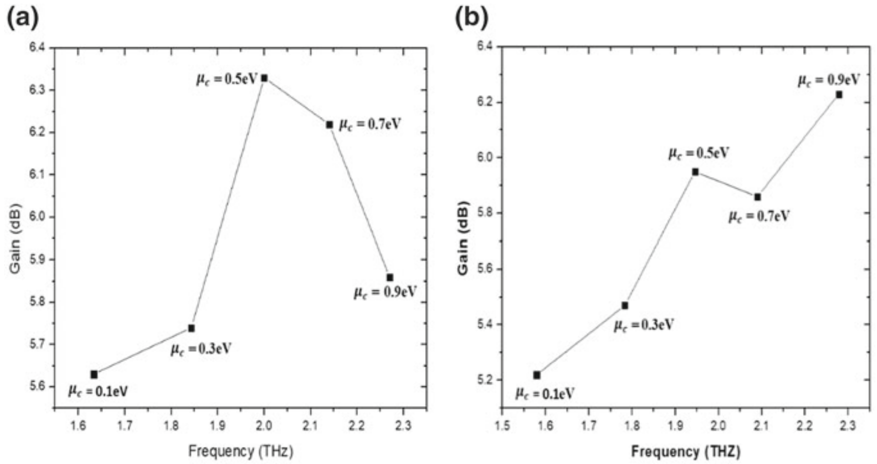


Fig. 4 Directive gain versus frequency curve for **a** simple patch antenna **b** U slot loaded patch antenna for different values of $\boldsymbol{\mu_c}$

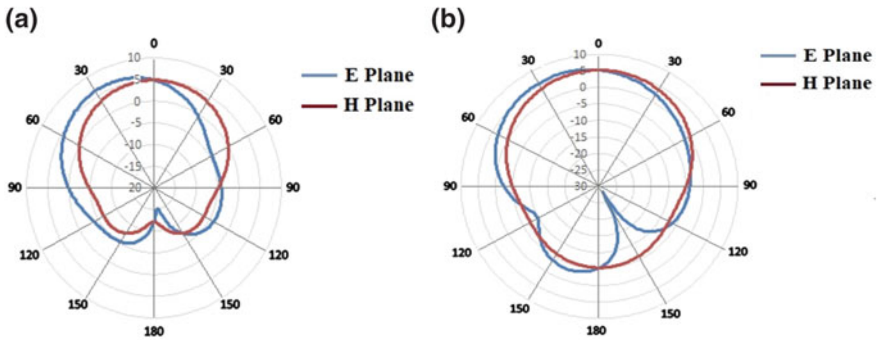


Fig. 5 Radiation pattern for **a** simple patch antenna **b** U slot loaded patch antenna for $\boldsymbol{\mu_c} = 0.5$ eV

effect. A simple patch antenna and U slot loaded patch antenna has been designed at THz frequencies. The simulated results of these antennas for different values of chemical potential are presented in this paper. It has been shown that the bandwidth of the Microstrip antenna can be increased using U slot loaded patch (Fig. 3).

References

1. Geim, A.K., Novoselov, K.S.: The rise of graphene. Nat. Mater. **6**(3), 183–191 (2007)
2. Castro Neto, A.H., et al.: The electronic properties of graphene. Rev. Modern Phys. **81**(1), 109–162 (2009)

3. Wu, Y.H., Yu, T., Shen, Z.X.: Two-dimensional carbon nanostructures: Fundamental properties, synthesis, characterization, and potential applications. J. Appl. Phys. **108**(7), 071301 (2010)
4. Yurchenko, S.O., Komarov, K.A., Pustovoit, V.I.: Multilayer-graphene-based amplifier of surface acoustic waves. AIP Adv. **5**(057144), 1–12 (2015)
5. Al-Dirini, F., Mohammed, M.A., Hossain, F.M., Nirmalathas, T., Skafidas, E.: All-graphene planar double-quantum-dot resonant tunneling diodes. J. Electron Device Soc. **4**, 30–39 (2016)
6. Fahad, M.S., Srivastava, A., Sharma, A.K., Mayberry, C.: Analytical current transport modeling of graphene nanoribbon tunnel field-effect transistors for digital circuit design. IEEE Trans. Nanotechnol. **15**, 39–50 (2016)
7. Li, Y., et al.: Graphene-based floating-gate nonvolatile optical switch. IEEE Photon. Technol. Lett. **28**, 284–287 (2016)
8. Mao, X., et al.: Optoelectronic mixer based on graphene FET. IEEE Electron Device Lett. **36**, 253–255 (2015)
9. Correas-Serrano, D., Gomez-Diaz, J.S., Perruisseau-Carrier, J., Alvarez-Melcon, A.: Graphene-based plasmonic tunable low-pass filters in the terahertz band. IEEE Trans. Nanotechnol. **13**, 1145–1153 (2014)
10. Wang, X.-C., Zhao, W.-S., Hu, J., Yin, W.-Y.: Reconfigurable terahertz leaky-wave antenna using graphene-based high-impedance Surface. IEEE Trans. Nanotechnol. **14**, 62–69 (2015)
11. Zangeneh-Nejad, F., Safian, R.: A tunable high-impedance THz antenna array. In: 23rd Iranian Conference on Electrical Engineering (2015)
12. Yao, G., et al.: Dynamically electrically tunable broadband absorber based on graphene analog of electromagnetically induced transparency. IEEE Photon. J. **8**(7800808), 1–8 (2016)
13. Wang, Y., Chen, Q., Shen, X.: Actively controlled plasmonic Bragg reflector based on a graphene parallel-plate waveguide. AIP Adv. **5**(077152), 1–7 (2015)
14. Zheng, R., Gao, D., Dong, J.: Ultra-compact broadband tunable graphene plasmonic multimode interferometer. IEEE Photon. Technol. Lett. **28**, 645–648 (2016)
15. Huang, C.-H., Yu, S.-C., Lai, Y.-C., Chi, G.-C., Yu, P.: Efficiency enhancement of organic/GaAs hybrid photovoltaic cells using transparent graphene as front electrode. IEEE J. Photovolt. **6**, 480–485 (2016)
16. de Oliveira, R.E.P., de Matos, C.J.S.: Graphene based waveguide Polarizers: In Depth physical analysis and relevant parameters. Graphene and Nanomaterials Research Center, Mackenzie Presbyterian University, Sao Paulo, 01302-907, Brazil (2015)
17. Conteduca, D., Dell'Olio, F., Ciminelli, C., Armenise, M.N.: Resonant graphene-based tunable optical delay line. IEEE Photonics J. **7**(7802409), 1–9 (2015)
18. Hanson, G.W.: Dyadic greens functions and guided surface waves for a surface conductivity model of graphene. J. Appl. Phys. **103**, 064302 (2008)

Concurrent Dual-Band Double-Layer High Gain Planar Antenna for WAICs/ITS Application

Shivesh Tripathi, Nagendra Prasad Pathak and M. Parida

Abstract Safety is the prime concern which drives the development of vehicular communication as support to intelligent transportation systems. The communication networks play a key role and supported by different components among which antenna plays a very crucial role. This paper presents a compact and new design of circular polarized (CP) patch antenna for use in wireless avionics intra-communications (WAICs) and intelligent transportation systems (ITS). Experimental results of the reflection coefficient, gain, and the radiation pattern is consistent with simulation. The prototype antenna can operate in a 4.4 GHz WAICs band with reflection coefficient under −15 dB and gain 8.3 dBi. In 5.9 GHz ITS band it operates with reflection coefficient under −15 dB and gain 8.8 dBi.

Keywords Double-layer · Dual-band · Wireless avionics intra-communications
Intelligent transportation systems · Beam width · Unidirectional · Coaxial-feed

1 Introduction

Wireless avionics intra-communication (WAICs) have an enormous potential to enhance an aircraft performance through more practical flight operation, reduction in the overall weight, reduction in the maintenance cost, and improvement of the safety [1]. The utilization of wireless technologies in aircraft lessens the measure of fuel required to fly by 12% through the weight decrease. United nation agency voted

S. Tripathi (✉) · N. P. Pathak · M. Parida
Indian Institute of Technology Roorkee, Roorkee 247667, Uttarakhand, India
e-mail: shiveshtripathi@gmail.com

N. P. Pathak
e-mail: nagppfec@gmail.com

M. Parida
e-mail: mparida@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_5

to grant a frequency band from 4.2 to 4.4 GHz for WAIC system to allow for the heavy wiring utilized in aircraft to be substituted by wireless networks [1].

Transport is a prime aspect of human life and with the help of intelligent transportation systems (ITS) which is the totally automated scenario of transportation system. The radio waves play a vital role in the development of ITS vehicle location and tracking systems. A wireless communication service called Dedicated Short-range Communications (DSRC) has been developed for vehicular communication can also be utilized for vehicle tracking. Federal Communications Commission (FCC) has allocated 75 MHz (5850–5925 MHz) bandwidth at 5.9 GHz for DSRC applications. In Europe, a frequency band at 5.9 GHz has been allocated to safety and non-safety DSRC applications [2]. DSRC is supposed to deliver a set of rapidly changing information, which can be both location and time dependent. DSRC systems are designed for the vehicle-to-vehicle and vehicle-to-infrastructure communication environments using a short-range wireless link for ITS applications. Broad band wireless systems play an increasingly important role by providing high-speed wireless links between many ITS subsystems [3].

Wireless link based tracking will improve safety and decrease traffic congestion. Motivated by above problem various antenna solutions are available in the literature which can be used for ITS and WAICs applications. Smart antennas can greatly enhance the performance of wireless systems and fulfill the requirement of improving coverage range, capacity, data rate and quality of service. Responsibility lies with the ITS designer to understand the working of a particular smart antenna before it is used for the intended operating environment [3]. Microstrip antennas are the preferred choice in modern mobile communications, in addition, these are easier to design, light weight, low coast and can be easily fabricated.

A comprehensive performance comparison has been carried out of industrial wireless networks for WAICs [4]. In [5] the design issues and current trends for WAIC has been discussed which paves the way for future WAIC. In [6] the measurement approach of interference path loss between the WAIC systems and aircraft system has been discussed. An antenna array is designed which is circularly polarized and operates on C-band, the antenna is compact and single layer which minimize the overall dimension without sacrificing the bandwidth presented in [7]. In [8] the measurement of parameters like safety and traffic of the mobile vehicle to vehicle propagation channel has been carried out at 5.9 GHz. The design of a low profile antenna for dual-band application is reported in [9]. Different antenna architecture for frequency bands like WLAN/WiMAX, DSRC, and GPS has been reported in the context of vehicular and general communication has been reported [10–14]. In [15–23] front end for DSRC and ITS application such as different antenna architecture with dual-band focusing on DSRC application has been reported. The return loss in such cases is more in both bands as well as the gain is minimal. To overcome

Table 1 Dimensions of the proposed antenna

Parameter	l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9	l_{10}	l_{11}	l_{12}	h_1
Value (mm)	9.5	11.5	7.63	9.5	9.5	9.19	3.5	4	2	3	11.5	11.5	1.28
Parameter	l_{13}	l_{14}	l_{15}	l_{16}	l_{17}	l_{18}	w_1	w_2	w_3	w_4	L	W	
Value (mm)	10.606	11.5	11.5	10.606	10	8	0.5	12.5	0.5	6.5	32	32	

these challenges a novel antenna with −15 dB return loss in both the bands with a high gain and broadside radiation pattern has been designed. To cover the region near a vehicle in all direction the antenna should be polarized circularly. Inter-vehicle communication (IVC) systems (i.e., systems not relying on road-side infrastructure) have the potential to radically improve the safety, efficiency, and comfort of everyday road travel is reported in [24]. Design and development of a planar dual-band GPS and DSRC antenna for road vehicles has been described in [25].

This paper reports design, analysis and characterization of a concurrent Dual-Band Double-Layer Planar Antenna for WAICs/ITS Application. Section 2 of this paper describes the geometry of the prototype antenna. Parametric analysis has been carried out to optimize the dimensions of proposed antenna in Sect. 3. Section 4 presents the detailed simulation and experimental characterization of the antenna with comparative study available in literature followed by a conclusion.

2 Design and Geometry of Proposed Planar Antenna

The geometry of the proposed antenna is shown in Fig. 1. The antenna is designed on low profile Rogers R3010 substrate with dielectric constant of 10.2, the material thickness of 1.28 mm. A 50 Ω SMA coaxial probe is used as the feeding structure to feed the antenna. The initial parameters of the antenna are derived as given in Balanis [23]. The dimensions are calculated above then optimized to achieve the desirable results. The Ansoft High-Frequency Structural Simulator (HFSS) has been used in the full wave simulation of proposed antenna was used to design and analyze. The design of the proposed antenna has been verified by simulation and measurement characteristics. The simulated characteristic the optimized parameter dimensions of the proposed dual-band double-layer planar antenna are given in Table 1.

The length and width of the proposed antenna are $L \times W$ which is 32 × 32 mm. The length of the corner truncation, l_2 and l_4 is given as 11.5 mm and 9.5 mm respectively. The feeding provided to the antenna by using a coaxial cable probe below the center of the patch, by determining the optimum feed location for maximum radiation. The bottom patch slot parameters are given by w_1, l_7, l_8 given in Fig. 1a. Upper patch slot is determined by w_3, l_{17}, l_{18} and given in Fig. 1b. The feed slot of the proposed

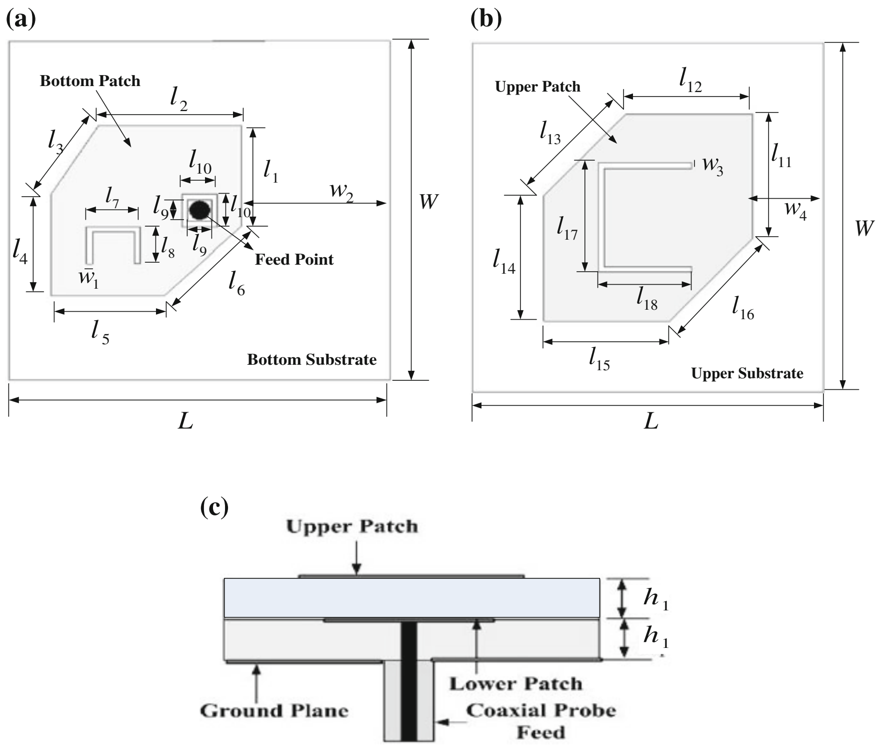


Fig. 1 Schematic structure of dual-band double-layer planar antenna. **a** Bottom patch antenna geometry. **b** Upper patch antenna geometry. **c** Complete antenna geometry

antenna is given by l_9 and l_{10}. The proposed dual-band, the double-layer antenna prototype is fabricated and shown in Fig. 2. For better understanding the effects of key parameters on the antenna performance a parametric investigation is performed which is discussed in the following section.

3 Results and Discussion

The proposed antenna is designed and fabricated to validate against simulated design. The return loss characteristic of the proposed antenna is measured using vector network analyzer. The measured and simulated return loss is shown in Fig. 3a. The simulated results reveal the bandwidth of the antenna is 70 MHz (4.38–4.45 GHz) in the first operating band and 200 MHz (5.78–5.98) in the second operating band. The

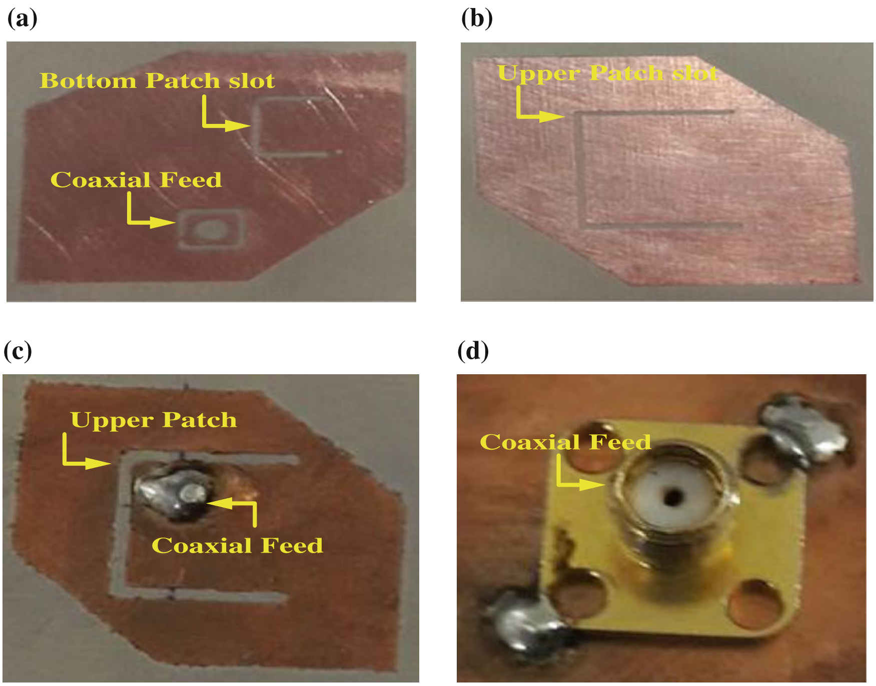


Fig. 2 Fabricated prototype of proposed antenna. **a** Bottom patch. **b** Upper patch. **c** Upper and bottom patch combined. **d** Coaxial feed connector

measured bandwidth is found to be ~75 MHz (4.38–4.55 GHz) in the lower frequency band and 200 MHz (5.82–6.02 GHz) in the upper-frequency band. These operating bands of the proposed antenna make it suitable for the operation in 4.4 GHz WAICs and 5.9 GHz ITS band applications. As per the measured results, the proposed antenna has better than −15 dB of the reflection coefficient at WAICs (4.38–4.55 GHz) and ITS (5.850–5.925 GHz) bands respectively. Figure 3b shows the simulated and measured antenna gain of the proposed antenna. The maximum measured gain of the antenna is 8.3 and 8.8 dBi at 4.4 GHz and 5.9 GHz, respectively.

Figure 4 shows the simulated and measured characteristics of the far-field radiation patterns (X-Z plane and Y-Z plane) of proposed antenna at 4.4 GHz, respectively. As shown in Fig. 4 the results show patterns are broadside, therefore the proposed antenna is a good candidate for dual-band broad side pattern in nature. Figure 5 shows the simulated and measured characteristics of the far-field radiation patterns (X-Z plane and Y-Z plane) of proposed antenna at 5.9 GHz, respectively. The results show that the proposed antenna is broadside radiation pattern. Figure 6 shows the surface current distribution over the plane of the designed antenna. The current

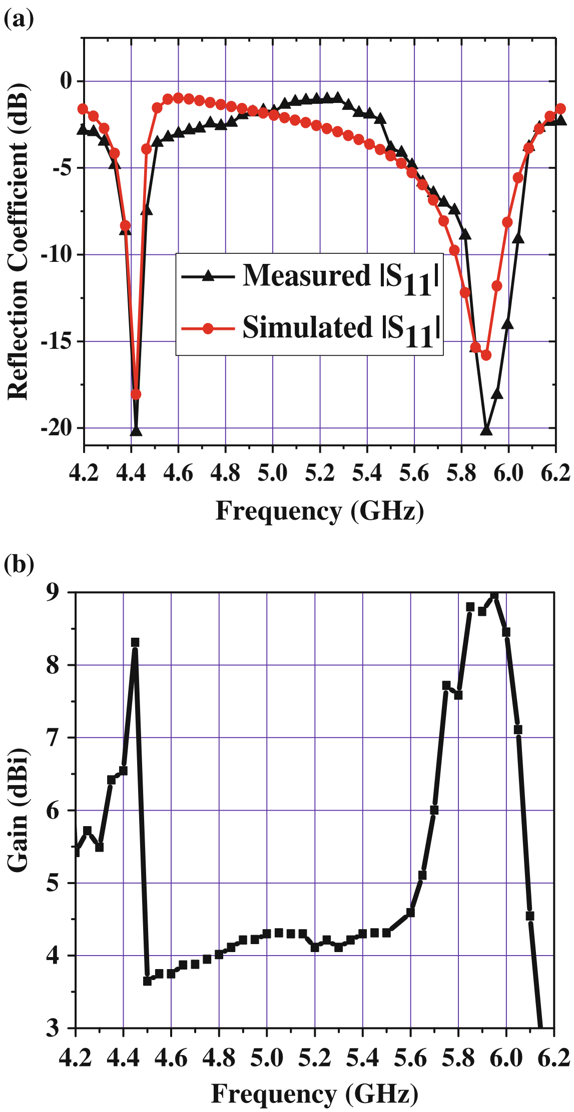


Fig. 3 **a** Simulation and measurement results for reflection a coefficient. **b** Maximum antenna gain

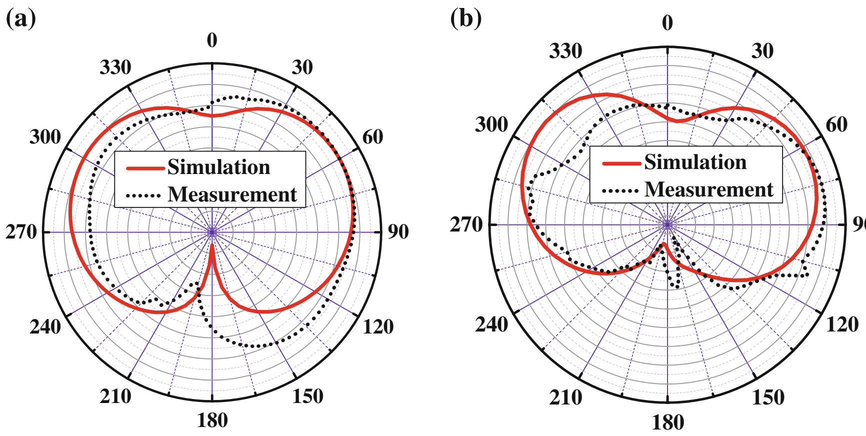


Fig. 4 Simulated (solid line) and measured (dotted line) results of the designed antenna radiation pattern at 4.4 GHz. **a** X-Z plane radiation pattern at 4.4 GHz. **b** Y-Z plane radiation pattern at 4.4 GHz

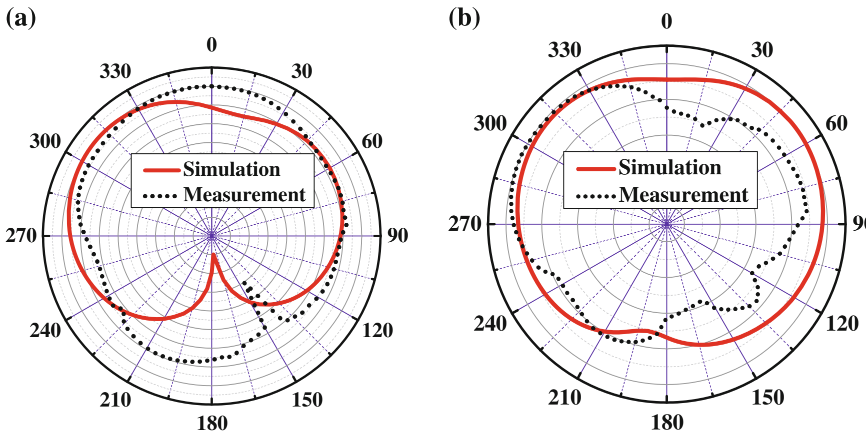


Fig. 5 Simulated (solid line) and measured (dotted line) results of the designed antenna radiation pattern at 5.9 GHz. **a** X-Z plane radiation pattern at 5.9 GHz. **b** Y-Z plane radiation pattern at 5.9 GHz

distribution on the feed line, ground plane, and patch for dual resonant frequencies are shown in Fig. 6 the flow of the current at the bottom slot is opposite to the flow at the corner of patch (Fig. 6a). The energies coupled between them produce 4.4 GHz resonant frequency. Similarly, 5.9 GHz resonant frequency is produced due to energy coupling between the feed line and patch (Fig. 6b). The flows of current between the slots are found to be in the opposite directions. The energies coupled between them produce 4.4 and 5.9. Figure 7 shows the 3D polar characteristics of the designed antenna at 4.4 GHz and 5.9 GHz, respectively. As per the simulation

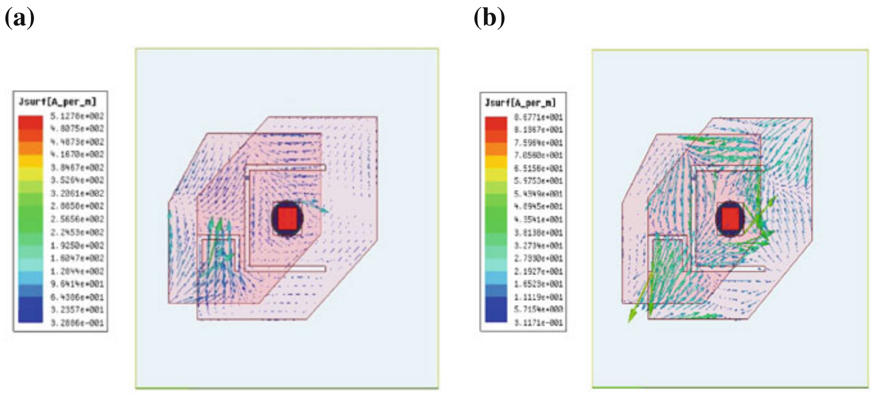


Fig. 6 Surface current distribution over the plane of the antenna. **a** 4.4 GHz and **b** 5.9 GHz

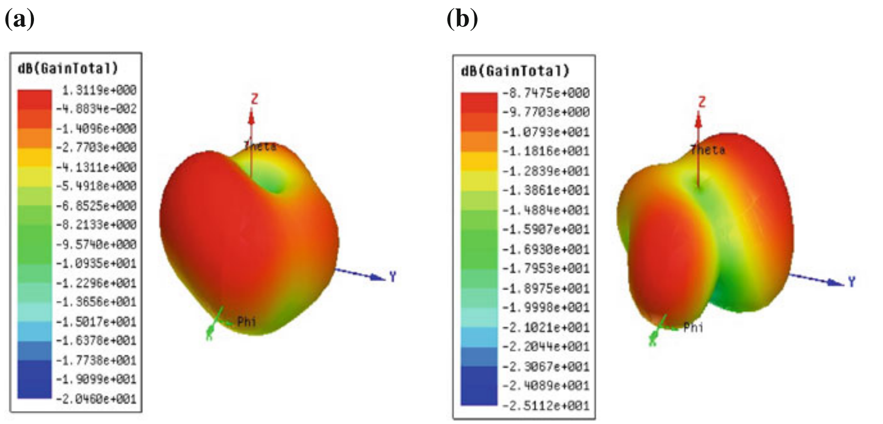


Fig. 7 Simulated 3 D polar characteristics of antenna radiation pattern of **a** 4.4 GHz and **b** 5.9 GHz

and measurement characteristics, the proposed planar dual-band double-layer patch antenna is a suitable candidate for WAICs/ITS applications.

The comparison of proposed antenna with other reported works available in the open literature is shown in Table 2.

Finally, a comparison of proposed work with recently published work on dual-band double-layer patch antenna is shown in Table 2. This table shows the proposed work gives a better result than the recently published work. Hence it is most suitable for dual-band double-layer patch antenna with a suitable for WAICs/ITS applications.

Table 2 Comparison of the proposed antenna with those available in literature

References	Antenna type	Frequency (GHz)	S_{11} (dB)	Gain (dBi)	Type/Nature of radiation pattern
[7]	Dual-band	1.575 and 5.88	−10 and −10	5 and 2.45	Omnidirectional pattern
[8]	Dual-band	2.4 and 5.7	−10 and −10	2.48 and 5.9	Unidirectional radiation pattern
[10]	Dual-band	5.1 and 5.95	−15 and −10	2 and 2.2	Omnidirectional pattern
[12]	Dual-band	1.57 and 5.88	−15 and −10	1.43 and 3.51	Linear radiation pattern
[15]	Dual-band	1.57 and 5.9	−10 and −10	2 and 1.5	Hemispherical radiation pattern
This work	Dual-band	4.4 and 5.9	−15 and −15	8. 3 and 8.8	Broad side pattern

4 Conclusion

A novel compact, dual-band double-layer antenna for dual-band operation (WAICs/ITS) with prescribed pattern and polarization has been designed and validated. The proposed planar antenna operates in 4.4 and 5.9 GHz band and having 75 MHz and 200 MHz BW; hence it is suitable for WAICs band (4.4 GHz) and ITS band (5.9 GHz) applications. The simulated and measured results show good radiation patterns in E-plane, and H-plane patterns and with measured antenna gain of 8.3 and 8.8 dBi at 4.4 and 5.9 GHz, respectively.

References

1. Technical characteristics and spectrum requirements of wireless avionics intra-communications systems to support safe operation Intonation Telecommunication Unit (ITU), genera, Switzerland document 2013ITU-RM2283-0
2. Tripathi, S., Pathak, N.P., Parida, M.: Design of a planar antenna for vehicle tracking using DSRC. In: Proceedings of the 3rd Conference of Transportation Research Group of India (3rd CTRG), pp. 1–10, Kolkata, India, 17–20 Dec 2015
3. Dhar, S., Kandar, D., Bose, T., Bera, R.: Smart antenna based broadband communication in intelligent transportation system (2009). arXiv:0903.3163
4. Park, P., Woohyuk, C.: Performance comparison of industrial wireless networks for wireless avionics intra-communications. IEEE Commun. Lett. **21**, 116–119 (2016)
5. Sámano-Robles, R., Eduardo, T., João, C., André, R.: Wireless avionics intra-communications: current trends and design issues. In: Eleventh International Conference on Digital Information

Management (ICDIM). IEEE, pp. 266–273 (2016)
6. Engel Brecht, J., Fuss, T., Schwark, U., Michler, O.: Measurement of interference path loss between wireless avionics intra-communications system and aircraft systems at 4.2–4.4 GHz band. In: Lough borough Antennas and Propagation Conference (LAPC), pp. 119–123 (2014)
7. Maddio, S.: Compact wideband circularly polarized antenna array for C-band applications. IEEE Antennas Wirel. Propag. Lett. **14**, 1081–1084 (2015)
8. Cheng, L., Benjamin, H.E., Daniel, S.D., Fan, B., Priyantha, M.: Mobile vehicle-to-vehicle narrow-band channel measurement and characterization of the 5.9 GHz dedicated short range communication (DSRC) frequency band. IEEE J. Sel. Areas Commun. **25**, 1501–1516 (2007)
9. Rafi, G.Z., Mohajer, M., Malarky, A., Mousavi, P., Safavi-Naeini, S.: Low-profile integrated microstrip antenna for GPS-DSRC application. IEEE Antennas Wirel. Propag. Lett. **8**, 44–48 (2009)
10. Dedicated short range communication (DSRC) for transport information and control systems (TICS), ARIB Standard, Nov 1997
11. Zhou, H., Hong, W, Tian, L., Cheng, L.: Cross-polarization suppressed and linearly polarized patch array antenna with cavity-backed slot feed for vehicle applications. IEEE Antennas Wirel. Propag. Lett. **15**, 126–129 (2016)
12. Laohapensaeng, T.: Design of dual-polarized microstrip antenna for ITS applications. In: International Conference on Digital Arts, Media and Technology (ICDAMT), pp. 43–47 (2017)
13. Khan, Q.U., Ihsan, M.B., Fazal, D., Malik, F.M.: Higher order modes: a solution for high gain, wide band patch antennas for different vehicular applications. IEEE Trans. Veh. Technol. **26**, 1–7 (2016)
14. Sim, C.Y.D., Yang, B.H.A.: Single layer dual-band CP microstrip antenna for GPS and DSRC applications. J. Electromagn. Waves Appl. **22**, 529–539 (2008)
15. Cidronali, A., Stefano, M., Giovanni, Gianfranco, M.: Design trade-off for a compact 5.8 GHz DSRC transponder front-end. Microw. Opt. Technol. Lett. **57**, 1187–1191 (2015)
16. Varum, T., Matos, J., Pinto, P., Abreu, R.: Non-uniform broadband circularly polarized antenna array for vehicular communications. IEEE Trans. Veh. Technol. **65**, 7219–7227 (2016)
17. Malarky, A., Rafi, G.Z., Naeini, S.S.: A planar dual band GPS and DSRC antenna for road vehicles VTC-2007. 2096–2100 (2007)
18. Dar, K, Bakhouya, M., Gaber, J., Wack, M.: Wireless communication technologies for ITS applications. IEEE Commun. Mag. **48**, 56–162 (2010)
19. Mondal, T., Samanta, S., Ghatak, R., Chaudhuri, S.B.: A novel circularly polarized DSRC band square microstrip antenna using minkowski fractal structure for vehicular communication. In: IEEE Conference on Vehicular Electronics and Safety (ICVES), pp. 141–146 (2014)
20. Leonardi, O., Pavone, M., Cadili, T., Sorbello, G., Isernia, T.: Monolithic patch antenna for dedicated short-range communications. Electron. Lett. **49**, 85–86 (2013)
21. Kim, S.M., Yang, W.G.: Single feed wideband circular polarized patch antenna. Electron. Lett. **43**, 703–704 (2007)
22. Kishore, N., Prakash, A., Tripathi, V.S.: A reconfigurable ultra-wide band antenna with defected ground structure for ITS application. AEU-Int. J. Electron. Commun. **72**, 210–215 (2017)
23. Balanis, A.: Antenna Theory Analysis and Design, 3rd edn. Wiley (2005)
24. Sichitiu, M.L., Kihl, M.: Inter-vehicle communication systems: a survey. IEEE Commun. Surv. Tutor. **10**, 88–105 (2008)
25. Malarky, A., Rafi, G.Z., Safavi-Naeini, S.: A planar dual band GPS and DSRC antenna for road vehicles. In: Proceedings of VT C-20 07 Fall. IEEE 66th Vehicular Technology Conference (2007)

Compact Rat-Race Coupler-Based Microstrip Balun Without Any Isolation Port

Ankita Kumari, Tamasi Moyra and Priyansha Bhowmik

Abstract In this work, a compact miniaturized microstrip balun is proposed by removing the isolation port of a Rat-Race Coupler (RRC). The proposed balun consists of six quarter-wavelength Transmission Line (TL), and the TL is designed using interdigital capacitor and high impedance TL in parallel configuration. The fractional bandwidth (FBW) of RRC and balun are 37.5% and 33%, respectively, at centre frequency 2.4 GHz for amplitude imbalance of ± 1 dB and phase imbalance of $\pm 8°$. The size occupied by the proposed balun is $0.51\ \lambda_g \times 0.22\ \lambda_g$.

Keywords Rat-race coupler · Balun · Miniaturized balun · Interdigital capacitor

1 Introduction

Rat-Race Coupler (RRC) and baluns are widely used in microwave communication circuit. RRC is a four-port device and consists of three λ/4 TLs and one 3λ/4 TL (formed by joining three quarter-wavelength TL). It is used to combine and divide the power, especially in RF systems on chip [1, 2]. Balun is a three-port device, which converts an unbalanced signal to balanced signal. It acts as a building block of mostly RF circuit such as single-ended power amplifier frequency multipliers, balanced mixers and so on.

The conventional RRC occupies a large area which becomes a limitation in practical microwave circuits. Many miniaturization techniques have been proposed to develop a compact RRC with wide bandwidth [3–12]. The conventional ring hybrid [3] has 23.6% fractional bandwidth (FBW) and occupies a large area. In [4], a modified ring hybrid of 12 dB power division has been designed using coupled line. The observed FBW is of 108% at the cost of large size. A 12 dB ring hybrid [5] has been designed by replacing two λ/4 line sections with coupled line section and open

A. Kumari (✉) · T. Moyra · P. Bhowmik
Department of Electronics and Communication Engineering, National Institute of Technology Agartala, Agartala, West Tripura 799046, India
e-mail: ankitakr.01@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_6

stubs. Though the size was miniaturized but the bandwidth reduced significantly. To reduce the size of RRC, the single TL is replaced by non-uniform TL [6], it gives a 3-dB coupling but provides 20% operational bandwidth. In [7–9], left-handed (LH) material is used which gives a wider bandwidth and a smaller size but the structure is multilayer which makes it costly. In [7], an artificial lumped section replaces the 3λ/4 TL, though it offered bandwidth enhancement but the size can be further reduced. A broadband RRC is designed using LH TL and multisection quarter wavelength [9], to widen the operational bandwidth but with a larger size. In [10], a reduced size RRC is proposed by replacing 3λ/4 TL with Lange couplers phase invertors but the bandwidth is reduced. A wideband RRC [11] has been designed with shorted coupled line section and multisection impedance transformer. This gives a wide bandwidth but the size is immense.

A novel compact wideband planer balun with high isolation is proposed [12], which is composed of a broadband coupled line power divider and a 180° phase shifter using coupled line and composite LH TL. It provides wide bandwidth but occupies a large area. A branch line balun with a quarter-wavelength stub and meandering a branch designed on Duroid 6006 which is costly and it provides 40% FBW but extensive size.

In this work, a compact balun with enhanced bandwidth has been achieved from an RRC. The proposed λ/4 TL section of impedance 70.7 Ω is achieved using interdigital capacitor and high impedance TL in parallel configuration. The ABCD parameter of the proposed TL is equalized to the conventional ABCD parameter to achieve the design equations. Based on the equations, an RRC operating at 2.4 GHz has been designed and simulated. On removal of the isolation port, a compact size balun is designed. The FBW of the balun is 33% and occupies 0.51 $\lambda_g \times 0.22\ \lambda_g$ areas.

2 Single Transmission Line Section

In this section, the design procedure of proposed quarter-wavelength TL has been described. Considering a single TL (Fig. 1), this holds some capacitive and inductive effect. An interdigital capacitor (two fingers of length l = 4 mm, width w = 0.8 mm and gap 1 mm between fingers) is connected in parallel with a high impedance TL. Interdigital capacitor gives a nonlinear phase response which is useful to reduce the dimension of TL. An equivalent circuit of proposed quarter-wavelength TL is shown in Fig. 2. The structure consists of one series inductance L_c with the Interdigital capacitance C_c and both are connected in parallel with an inductance L.

The characteristic impedance of the equivalent π-circuit can be given as

$$Z_c = \sqrt{\frac{L\left(j\omega L_c + \frac{1}{j\omega C_c}\right)}{C_p\left(j\omega L + j\omega L_c + \frac{1}{j\omega C_c}\right)}} \tag{1}$$

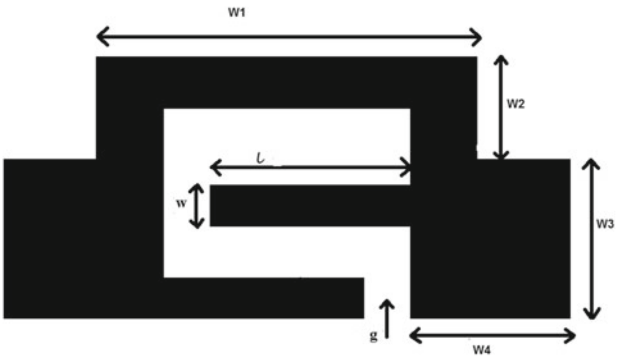


Fig. 1 Layout of single transmission line

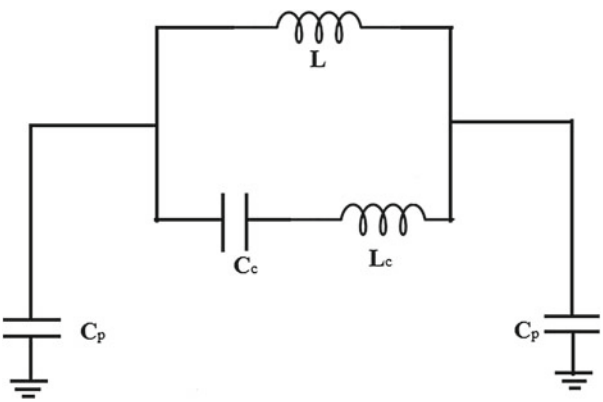


Fig. 2 Equivalent circuit of transmission line

As from the Eq. (1), one can easily figure it out that $Z_c \propto \sqrt{L}$ and $Z_c \propto 1/\sqrt{C_c}$. ABCD parameter of the half wavelength conventional TL is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 0 & jZc \\ j/Zc & 0 \end{bmatrix} \tag{2}$$

Considering the π-equivalent lumped circuit of proposed TL for calculation, the ABCD parameters are

$$A = \frac{1 - \omega^2 L_c C_c - \omega^2 L C_c - \omega^2 C_p L + \omega^4 L_c C_c C_p L}{1 - \omega^2 L_c C_c - \omega^2 L C_c} \tag{3a}$$

$$B = \frac{j\omega L\left(1 - \omega^2 L_c C_c\right)}{1 - \omega^2 L_c C_c - \omega^2 L C_c} \tag{3b}$$

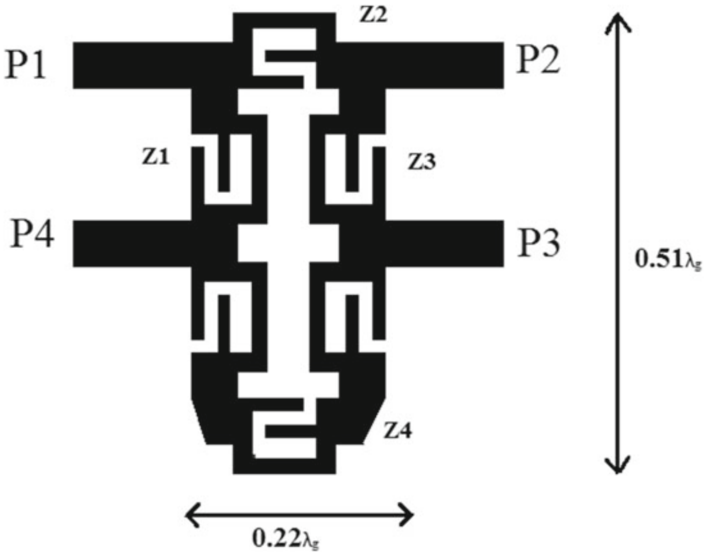


Fig. 3 Layout of RRC

$$C = \frac{j\omega\left(2C_p - 2\omega^2 L_c C_c C_p - 2\omega^2 C_p L C_c - \omega^2 C_p^2 L + \omega^4 C_p^2 L L_c C_c\right)}{1 - \omega^2 L_c C_c - \omega^2 L C_c} \tag{3c}$$

$$D = \frac{1 - \omega^2 L_c C_c - \omega^2 L C_c - \omega^2 C_p L + \omega^4 L_c C_c C_p L}{1 - \omega^2 L_c C_c - \omega^2 L C_c} \tag{3d}$$

After equating Eqs. (3a), (3b), (3c) and (3d) to the ideal value of ABCD parameter (2), we can get the value of $L_{c,}$ Cc, L and Cp, respectively.

Other physical measurements for the proposed TL (Fig. 2) are W1 = 6.8 mm, W2 = 2 mm, W3 = 3.1 mm, W4 = 2.9 mm, g = 0.8 mm.

3 Design and Discussion of RRC

The 3 dB RRC has been designed on FR4 substrate, with dielectric constant 4.4, substrate height 1.6 mm and loss tangent 0.02. To get a compact RRC, conventional TL is replaced by the proposed TL. Figure 3 shows a layout of the proposed RRC. It is simulated using Zeland IE3D-14 based on MoM's method.

Figure 4 shows the simulated S-parameter response of proposed RRC. The input signal splits equally between port 2 and port 4 at centre frequency 2.4 GHz. The isolation at port 3 and reflection from port 1 is −28 dB. The phase difference between port 2 and port 4 is 0° is shown in Fig. 5. The ±1 dB amplitude imbalance and ±8° phase tolerance ranges from 2 to 2.9 GHz.

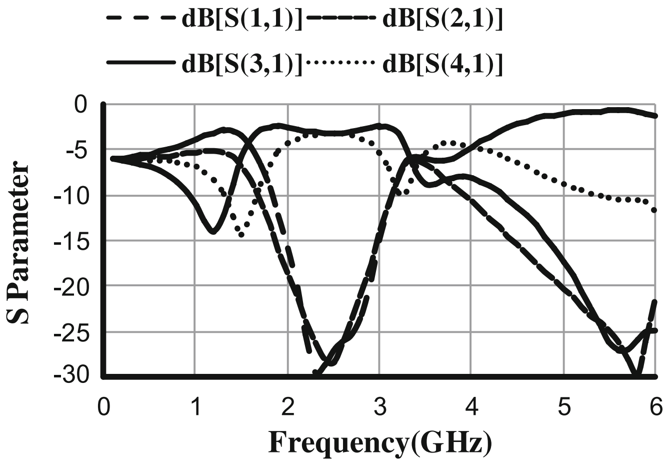
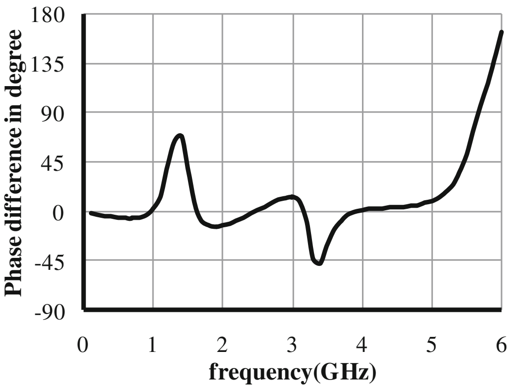


Fig. 4 Simulated S-parameter through port, coupling and isolation of RRC

Fig. 5 Phase difference between S21 and S41



4 Designs and Discussion of Balun

A balun has been designed by removal of isolation port of RRC, hence three ports are there, and two ports are used, to get the splitted input signal. To get the 180° phase difference between outputs, hence, the port P4 in Fig. 3 are adjusted as shown in Fig. 6 (proposed balun structure). It has been simulated in Zeland IE3D-14.

Figure 7 shows the simulated S-parameter response of balun. Figure 8 shows the phase difference of output at port 2 and port 3. The phase difference 180° and 3 dB bandwidth lies from 2 to 2.8 GHz with phase imbalance ±8° and amplitude imbalance of ±0.1 dB. The size of the proposed balun and RRC is 0.51 $\lambda_g \times 0.22$ λ_g ($\lambda_g = 58$ mm).

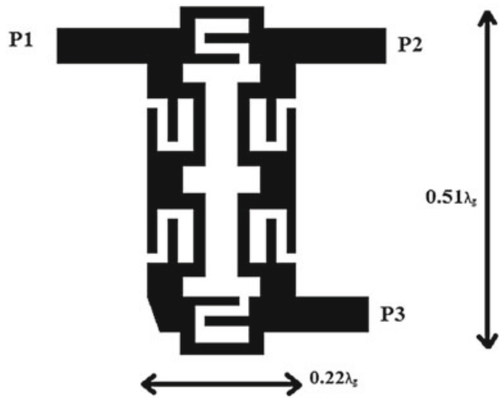


Fig. 6 Layout of the proposed balun

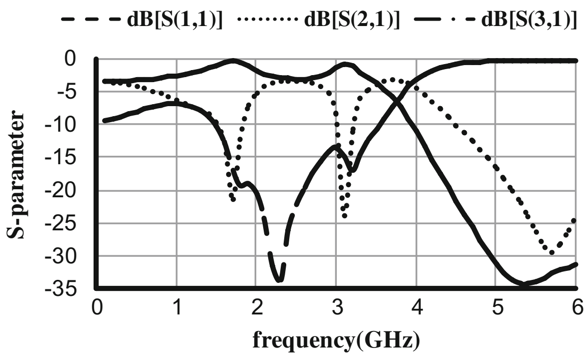


Fig. 7 Simulated S-parameter response of balun

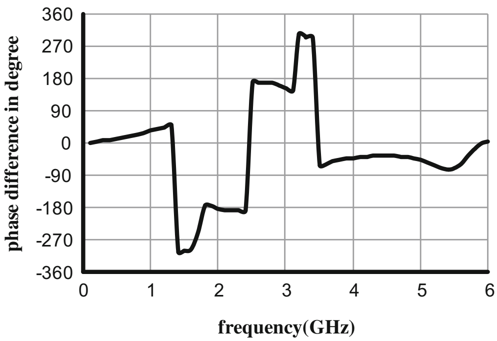


Fig. 8 Simulated phase difference of output at port 2 and port 3

5 Performance Comparisons

Reference	Substrate specification	Centre frequency (GHz)	FBW (%)	Size of balun (λ_g^2)	Size of balun (mm)
[12]	$\varepsilon_r = 3.48$, loss tangent = 0.0037	2.75	88.7	1.02	3431.28
[13]	$\varepsilon_r = 6.15$, loss tangent = 0.002	1.5	37.9	0.0585	252.81
This work	$\varepsilon_r = 4.4$, loss tangent = 0.02	2.4	33	0.1122	392.6

6 Conclusions

A compact balun is proposed and simulated to get 3 dB power division and 180° phase difference between outputs at centre frequency 2.4 GHz. The balun has been designed by removing the isolation port of RRC. RRC is a prototype to achieve a compact balun. The structure is compact, cheap and simple with better performance and can be used for practical microwave communication circuit.

References

1. Xu, H.X., Wang, G.M., Lu, K.: Microstrip rat-race couplers **12**(4), 117–129 (2011)
2. Iyer, B., Pathak, N., Ghosh, D.: Dual-input dual-output RF sensor for indoor human occupancy and position monitoring. IEEE Sens. J. **15**(7), 3959–3966 (2015)
3. Pon, C.Y.: Hybrid-ring directional coupler for arbitrary power divisions **9**(6), 529–535 (1961). MTT-9
4. Ahn, H.R., Nam, S.: Wideband microstrip coupled-line ring hybrids for high power-division ratios **61**(5), 1768–1780 (2013)
5. Ahn, H.R., Tentzeris, M.M.: A novel wideband compact microstrip coupled-line ring hybrid for arbitrarily high power-division ratios **64**(6), 630–634 (2017)
6. Bekasiewicz, A., Kurgan, P.: A compact microstrip rat-race coupler constituted by non-uniform transmission lines **56**, 970–974 (2014)
7. Okabe, H., Caloz, C., Itoh, T.: A compact enhanced-bandwidth hybrid ring using an artificial lumped-element left-handed transmission-line section **52**(3), 798–804 (2004)
8. Kim, T.G., Lee, B.: Metamaterial-based wideband rat-race hybrid coupler using slow wave lines **4**(6), 717–721 (2010)
9. Jakub, S., Ilona, P., Wincza, K., Gruszczynski, S.: Bandwidth improvement of rat-race couplers having left-handed transmission-line sections **24**, 341–347 (2014)

10. Liu, G.-Q., Wu, L.-S., Yin, W.-Y.: A compact microstrip rat-race coupler with modified lange and t-shaped arms **115**, 509–523 (2011)
11. Gruszczynski, S., Wincza, K.: Broadband rat-race couplers with coupled-line section and impedance transformers **22**(1), 22–24 (2012)
12. Wu, Y., Zhuang, Z., Jiao, L., Liu, Y.: A compact planar wide-band balun with high isolation based on coupled-line and composite right–left-handed transmission line **58**, 372–376 (2016)
13. Li, J.L., Qu, S.W., Xue, Q.: Miniaturised branch-line balun with bandwidth enhancement **43**(17), 931–932 (2007)

Application of the Fractal Defected Ground Structure in Design of the Bandpass Filter

Chandni V. Desai and Pravin R. Prajapati

Abstract The design of planar filter using stepped impedance type distributed synthesis has been proposed for 8.6–9.5 GHz frequency band. To improve the return loss characteristics, instead of using traditional defected ground structure (DGS), the combination of fractal geometry and DGS, known as 'fractal DGS' has been embedded in the ground plane of the filter. The filter using fractal geometry gives 65% better return loss as compared to filter without DGS. Three laboratory prototypes (without DGS, first iteration DGS and second iteration DGS) have been developed to validate the simulated and measured results. The measured results are agreed with simulated one.

Keywords Ground structure · Fractal geometry · Bandpass filter

1 Introduction

The concept of microwave filter is a fundamental part of the radio frequency frontend module in advanced wireless communication technique [1]. The filter is any kind of residual or active group with a specific frequency response in regarding to amplitude and phase. It can also be categorized, based on their applications; one of them is a bandpass filter. The bandpass filter is defined by the center frequency along with 3 dB bandwidth. Bandpass filter used for passing particular amount of frequency band and block all other frequencies in the stop band. For the better improvement in results of pass band, there are some advanced techniques are available. Here, Improvement in return loss has been considered for bandpass filter. Defected Ground Structures

C. V. Desai · P. R. Prajapati (✉)
Department of Electronics and Communication Engineering, Center of Research in Optical, Microwave and Antenna (CROMA), A. D. Patel Institute of Technology, Karamsad 388121, Gujarat, India
e-mail: pravinprajapati05@gmail.com; ec.pravin.prajapati@adit.ac.in

C. V. Desai
e-mail: 1505chandni@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_7

(DGS) and Fractal in Defected Ground Structures are included in these advanced techniques. The concept of DGS developed in recent years, primarily from the investigations of Photonic Band Gap (PBG) structures in electromagnetics. The PBGs, applied to electromagnetic (EM) uses, are now referred to as Electromagnetic Band Gap (EBG) structures [2]. DGS might be regarded as a simplified type of a printed EBG on a ground plane. DGS is comparatively an upgraded field of investigation as well as applications associated with printed circuits, microwave filters in addition to antennas. There are different shapes are available to create a DGS on the surface of microwave filter. DGS is a single defect or a small number of periodic defects on the ground plane of a microstrip antenna, which intrigued interests in recent years due to its attractive features such as the band-stop and slow wave effect [3, 4]. The DGS is explored for size reduction of the antenna, design approach for circular polarization, mutual coupling reduction in antenna arrays, and elimination of scan blindness in antenna arrays, to suppress higher harmonic modes in antennas, to improve radiation properties of an antenna [5–11], etc. DGS having an advantage if size reduction in filters as well as it reduces ripple component. The keyword fractal was founded by the French mathematician B. B. Mandelbrot in 1970, after his pioneering analysis on multiple naturally produced irregular and fragmented geometries not contained within the realms of standard geometry [12].

The Latin term fractus which is related to the verb fangere, it means to break. These types of geometries were usually discarded as formless, however Mandelbrot shows that specific amazing characteristics can be related to the. Many of those shapes were familiar well before him, but Mandelbrots analysis is most successful, because he discovered the general element in many of these irregular geometries and formulate his principle is depends on his conclusion [13] these types of these geometries are sub divisible with each division a version of the parent and this is the most important as well as unique ability with Euclidean geometry. Mandelbrot defines the word fractal in many ways. Thus, there depends mostly on the classification of their size of fractals.

The idea of fractal is functions on self-similarity as well as reduction of element. By the reason of the space-filling and self-similarity properties of the fractal shapes, the concept of fractal is already applied to the microwave engineering broadly to scale down the dimension and complexity of microwave components and used elements, such as EBG low-pass filters [14], branch-line couplers [15, 16], etc. There is different geometry for fractal is available.

In this paper, the application of fractal DGS in design and development of planar filter has been demonstrated. The combination of the fractal geometry and DGS gives more enhanced performance of slow wave effect [17] and impedance matching [18], which leads to improvement in roll off factor and return loss of the filter.

2 Geometry of the Proposed Filter

From the proposed methodology Fig. 1 shows the top view of a bandpass filter. Roger RO3010 material is considered as a substrate. The dielectric constant of this material is 10.2, thickness of 1.27 mm and loss tangent is 0.0022. The width of the design is 44.39 mm and length filter is 81.81 mm.

Figure 2 shows the bottom view of a bandpass filter. The rectangular shape DGS is embedded on the bottom side of the filter. The dimension of the DGS is 8 10 mm. The 1st iteration DGS adds additional lumped inductance and capacitance, and hence improvement in impedance matching has been accomplished. Figure 3 shows 2nd iteration DGS embedded on the ground plane of the planar filter. For getting improved return loss, fractal DGS plays a major role. Self-similar structure is added in rectangular shape DGS, which gives enhanced slow-wave effect and improved impedance matching.

Figure 4 shows the fabricated layout of bandpass filter, Fig. 5 shows the first iteration of the bandpass filter, which shows DGS structure. Figure 6 shows the physical structure of fractal in DGS.

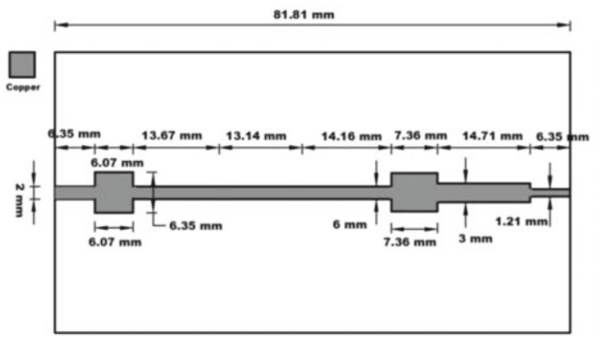


Fig. 1 Top view of a bandpass filter

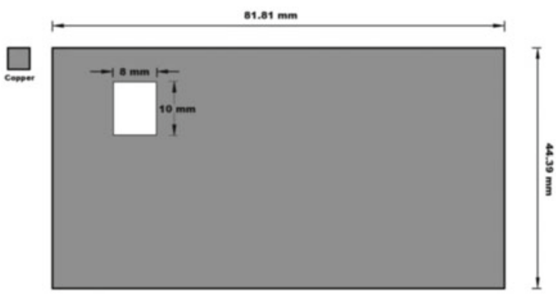


Fig. 2 Back view of bandpass filter with 1st iteration DGS

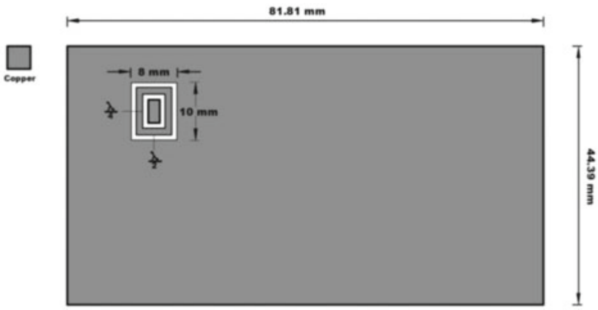


Fig. 3 Back view of bandpass filter with fractal in DGS

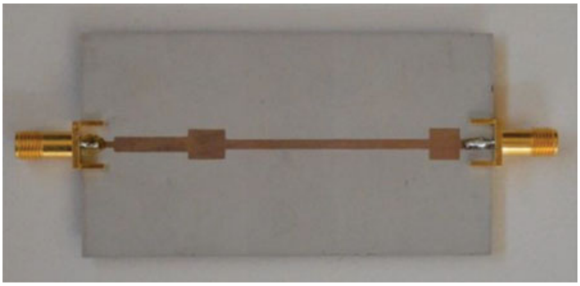


Fig. 4 Top view of fabricated layout of proposed model

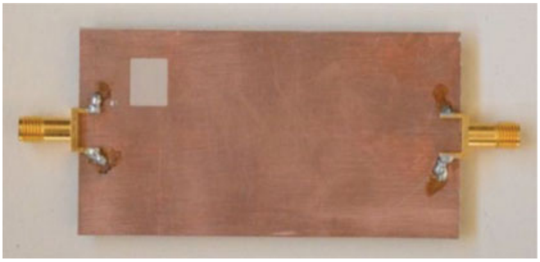


Fig. 5 Bottom view of fabricated layout of proposed model with 1st iteration DGS

Figure 4 shows the fabricated layout of bandpass filter, Fig. 5 shows the first iteration of the bandpass filter, which shows DGS structure. Figure 6 shows the physical structure of fractal in DGS.

Fig. 6 Bottom view of fabricated layout of proposed model with 2nd iteration fractal DGS

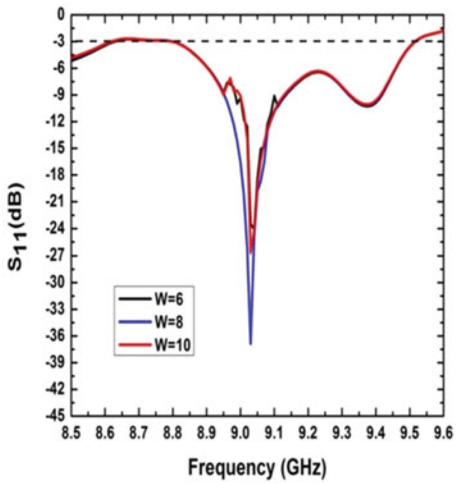


Fig. 7 Effect of variation of DGS width “W” on the return loss of the bandpass filter

3 Results and Discussion

To study the effect of variation of the dimensions of rectangular shape DGS, parametric sweep has been carried out using CST Microwave studio simulator. The overall width of the DGS (W) has been varied from 6 to 10 mm, with step of 2 mm.

As shown in Figs. 7 and 8, width of 8 mm gives optimum solution in terms of return loss and insertion loss.

Figure 9 and Fig. 10 show the effect of variation DGS length L on the return loss and insertion loss respectively. It is observed that length L = 10 mm gives optimum results, so L =10 mm has been considered for design. Figure 11 shows the simulated response of S11 parameter, which shows the return loss of bandpass filter for each frequency . The graph represents a different response of pass band using

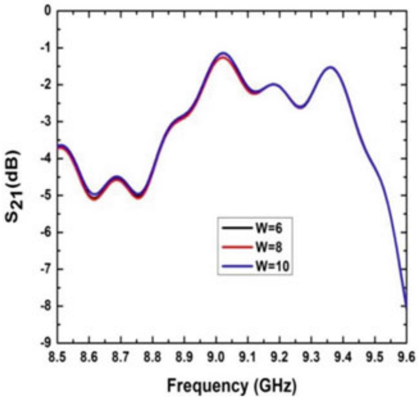


Fig. 8 Effect of variation of DGS width “W” on the insertion loss of the bandpass filter

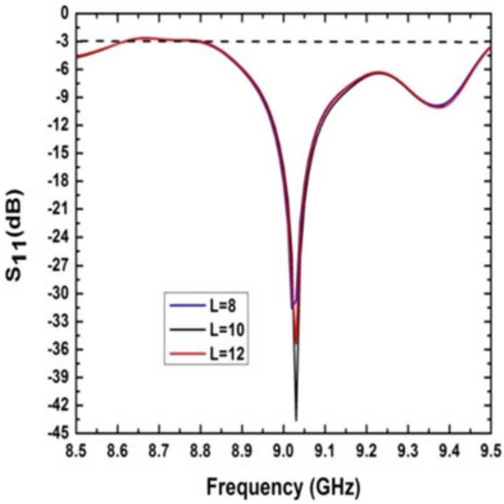


Fig. 9 Effect of variation of DGS width “L” on the return loss of the bandpass filter

three different techniques. The return loss of −26 dB, −33 dB and −43 dB has been achieved in case of without DGS, first iteration and second iteration fractal DGS respectively. It shows that, after embedding fractal DGS, the return loss improved by −26 dB to −42 dB, i.e., 65%. The filter passband frequency of 8.6–9.5 GHz.

Figure 12 shows the simulated response of S21 that means the insertion loss of the pass band. S21 parameter also shows the shape of the pass band. As shown in Fig. 12, all the three responses are almost overlapped with each other. So it is clear that in the proposed technique the insertion loss in each insertion remains the same which is −1.7 dB at 9.02 GHz frequency. It is concluded that there is no any adverse

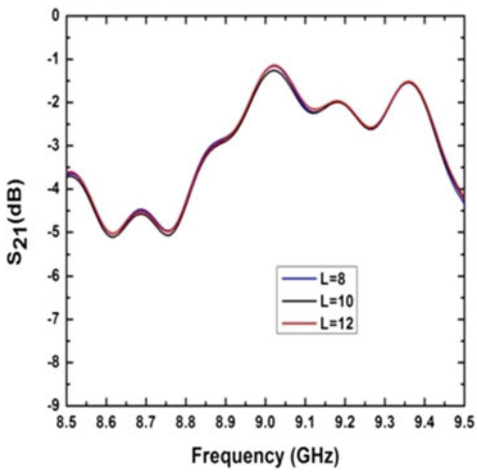


Fig. 10 Effect of variation of DGS width "L" on the insertion loss of the bandpass filter

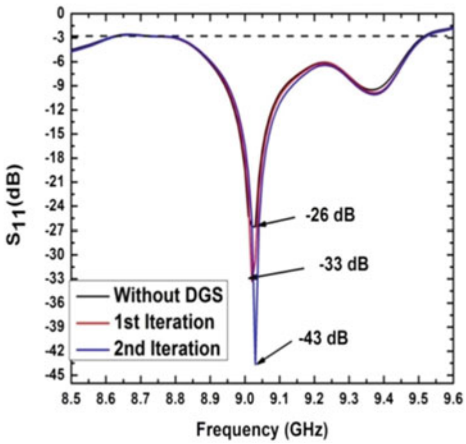


Fig. 11 Simulated S_{11} parameter of bandpass filter: without DGS, 1st and 2nd iteration

effect of embedding fractal DGS on insertion loss of the filter. The Propose technique has better return loss (−43 dB) in passband of bandpass filter as compared to [19] and [20], where load source coupling method and tradition DGS techniques are used and −28 dB and −29 dB return loss achieved respectively.

4 Conclusion

The application of Fractal Geometry and DGS in design and development of the planar filter has been demonstrated. The effects of variation of the DGS dimensions on the return loss and insertion loss have been demonstrated. It is seen that, after embedding the fractal DGS on the ground plane of the planar filter, the return loss

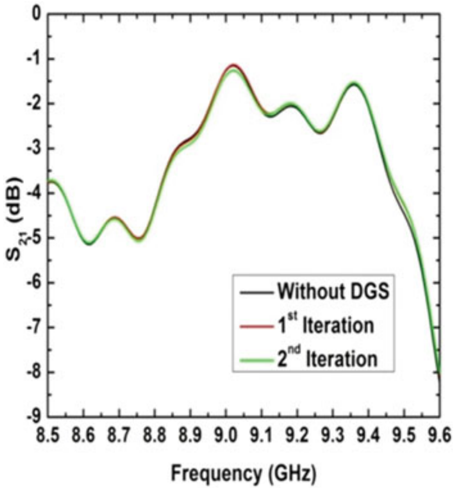


Fig. 12 Simulated parameter S_{21} of bandpass filter: without DGS, 1st and 2nd iteration

has been improved by 65% without any adverse effect on the insertion loss in the pass band of the filter. The study of third iteration fractal DGS and effect of different shape of the fractal DGS, to improve the insertion loss, roll off factor, selectivity, etc., may be the future scope of this work.

References

1. Khalida, S., Shafique, M.F.: Exact synthesis design theory of analogue broadband bandpass filter. Microelectron. J. **64**, 53–59 (2017)
2. Rahmat-Samii, Y., Mosallaei, H.: Electromagnetic band-gap structures: classification, characterization and applications. In: Proceedings of IEE-ICAP Symposium, vol. 2, pp. 5601–5644 (2001)
3. Iyer, B., Pathak, N., Ghosh, D.: Concurrent dualband patch antenna array for non-invasive human vital sign detection application. In: IEEE Asia-Pacific Conference on Applied Electromagnetics (APACE), pp. 150–153 (2014)
4. Arya, A.K.: Design studies of specific microstrip antennas with defected ground structure. Ph.D. thesis, Indian Institute of Technology Roorkee, Roorkee, India (2012)
5. Bark, H., Tavakoli, A.: Applications of spiral defected ground structures in design of a compact microstrip slot antenna. In: Proceedings of IEEE International Symposium on Antenna and Propagation Society, Honolulu, USA (2007)
6. Arya, A.K., Kartikeyan, M.V., Patnaik, A.: On the size reduction of microstrip antenna with DGS. In: Proceedings of 35th International Conference on Infrared Millimeter and Terahertz Waves (IRMMWTHz-2010), vol. 13, Rome, Italy (2010)
7. Thakur, J.P., Park, J.: An advanced design approach for circular polarization of microstrip antenna with unbalance DGS feed-line. IEEE Antennas Wirel. Propag. Lett. **5**, 101–103 (2006)
8. Salehi, M., Tawakoli, A.: A novel low mutual coupling microstrip antenna array design using detected ground structure. Int. J. Electron. Commun. **60**, 718–723 (2006)

9. Hou, D.B., Xiao, S., Wang, B.Z., Ziang, L., Wang, J., Hong, W.: Elimination of scan blindness with compact defected ground structures in microstrip array, antennas. IET Microw. Antennas Propag. **3**, 269–275 (2009)
10. Prajapati, P.R.: Application of defected ground structure to suppress out of band harmonics for WLAN microstrip antenna. Int. J. Microw. Sci. Technol. 1–9 (2015). Hindawi Publication
11. Zulkifli, F.Y., Rahardjo, E.T., Hartanto, D.: Radiation properties enhancement of triangular patch microstrip antenna array using hexagonal defected ground structure. Prog. Electromagn. Res. **5**, 101–109 (2008)
12. Peitgen, H.O., Jurgens, H., Saupe, D.: Chaos and Fractals: New Frontiers of Science. Springer, New York (1992)
13. Falconer, K.J.: Fractal Geometry: Mathematical Foundations and Applications. Wiley, New York (1990)
14. Chen, W.L., Wang, G.-M., Qi, Y.-N.: A compact wide-stop band Koch-shaped electromagnetic band gap microstrip low-pass filter. Microw. J. **10**(50), 160–166 (2007)
15. Chen, W.L., Wang, G.M.: Design of novel fractal-shaped branch-line couplers. Microw. Opt. Technol. Lett. **5**(50), 1198–1201 (2008)
16. Chen, W.L., Wang, G.M.: Exact design of novel miniaturized fractal-shaped branch-line couplers using phase-equalizing method
17. Kim, H.M., Lee, B.: Bandgap and slow/fast wave characteristics of defected ground structures including left-handed features. IEEE Trans. Microw. Theory Tech. **7**(54), 3113–3120 (2006)
18. Chen, Y., Manteuffel, D.: Distributed MEMS impedance matching network using a SWDGS technique. In: Loughborough Antennas & Propagation Conference (LAPC), pp. 1–4, Loughborough (2012)
19. Cao, L., Hu, J., Yin, L.: Compact coaxial dielectric bandpass filter with load-source coupling. In: Asia-Pacific Microwave Conference (APMC), pp. 1–3, Nanjing (2015)
20. Pal, B., Dwari, S.: A compact parallel coupled wideband bandpass filter with DGS and spurline. In: International Conference on Microwave and Photonics (ICMAP), pp. 1–2, Dhanbad, India (2015)

Design of UWB Monopole Antenna with Enhanced Gain Using Partially Reflective Surface

Pravin R. Prajapati and Shailesh B. Khant

Abstract The A low cost, high gain microstrip line feed UWB monopole antenna embedded with partially reflective surface is proposed in this paper. The antenna structure consists of UWB antenna, which acts as a main radiating element and that is fed with a array of total 30 square copper cells, which are considered on low permittivity substrate and suspended in air with the help of dielectric rods at height of 0.5 λ_0. The antenna with partially reflective surface gives enhancement of 1–2.7 dB gain in UWB range. The proposed structure is an attractive solution of poor gain problem in ultra wideband communication systems.

Keywords Partially reflective surface · UWB antenna · Monopole antenna

1 Introduction

Recently great attention is given to the design, development and research in printed ultrawide band antennas, because of easy fabrication, compact size and greater ability to integrate with other components that are fabricated with same printed circuit board. The features of patch antennae such as low cost, ease in fabrication and reproduction made it attractive to be used in RF systems on chip [1–3]. Several UWB monopole antennas have been reported in open literature, such as cylindrical conformal, circular disk, Octagonal-Shaped, rectangular, moon shaped, tapered microstrip slot, holly-leaf-shaped, semi-circle, triangular, two steps tapered [4–13]. Generally microstrip line feed monopole UWB antenna suffers from poor gain characteristics because

P. R. Prajapati (✉) · S. B. Khant
Department of Electronics and Communication Engineering, Center of Research in Optical, Microwave and Antenna (CROMA), A. D. Patel Institute of Technology, Anand 388121, Gujarat, India
e-mail: pravinprajapati05@gmail.com; ec.pravin.prajapati@adit.ac.in

S. B. Khant
e-mail: s_khnat@yahoo.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_8

of truncated ground. To enhance the gain of the microstrip line feed UWB planar antenna, partially reflecting surface (PRS) integrated antenna has been proposed in this paper.

2 Proposed Antenna Geometry and Design Guidelines

2.1 Design of the Proposed Antenna

Figure 1 shows the complete design dimension of the proposed microstrip line feed UWB monopole antenna. The Rogers 5880 material thickness (h) = 1.6 mm, dieletric constant (ε_r) = 2.2, loss tangent (tan $\delta = 0.0004$) of size = 20×40 mm^2 was considered as substrates for both UWB base antenna as well as PRS. The antenna is fed through a microstrip line feed arrangement using a 50 Ω sub-miniature type-A (SMA) connector. The antenna design was simulated and optimized using CST Microwave studio V. 17 simulator [14].

2.2 Design of Partially Reflected Surface

Thirteen square parasitic patches having dimensions 6×6 mm^2 arranged with six rows and five columns are considered on the lower side of substrate layer. The gap between each parasitic square element is 3 mm. The square parasitic patch dimensions and spacing between them are optimized in order to enhance the gain of the antenna. The length and width of each cell dimensions (d × d) are 7×7 mm^2, and the spacing or lattice constants of the patches is 3 mm as shown in Fig. 2a. In the proposed structure, the air is act as a dielectric medium between the main radiating element of the antenna and PRS to achieve high efficiency and wide bandwidth. The antenna is designed to operate over 3.8–8.8 GHz of ultra wide substrate are 53 mm and 63 mm respectively and thickness is 1.6 mm band. The substrate layer is positioned at a height of 0.5 λ_0 above the ground plane, where λ_0 is the free space wavelength corresponding to a central frequency of 6.5 GHz.

3 Results and Discussions

Figure 2b shows that bandwidth of 4.71 GHz (3.65–8.36 GHz) obtained with PRS, which is almost same as compared to without PRS, means there is no any adverse effect of PRS on the bandwidth of the antenna. The UWB antenna with PRS combine structure considered as a cavity resonator. Multiple reflections are done by the PRS towards the antenna. The distance between a radiating element and PRS is kept such

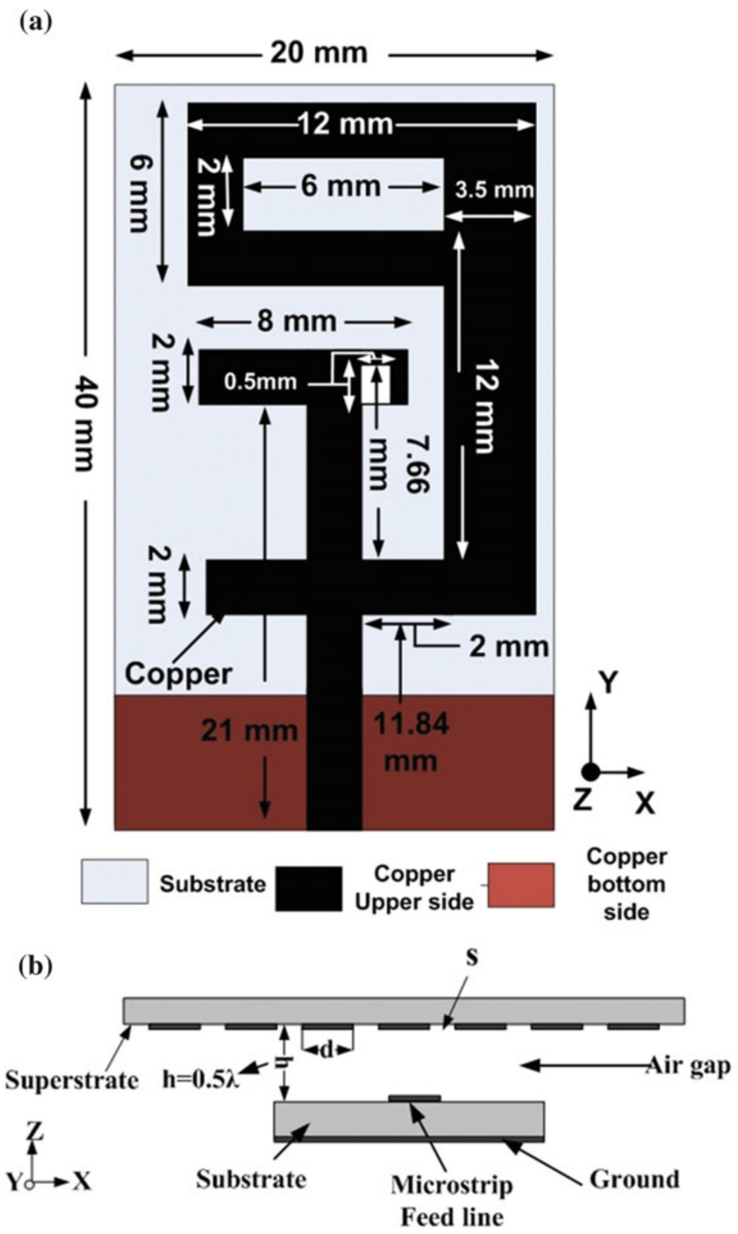


Fig. 1 Design geometry of UWB monopole antenna: **a** top view (without PRS) **b** front view of proposed antenna with PRS (d = 7 mm, s = 3 mm, h = 23 mm)

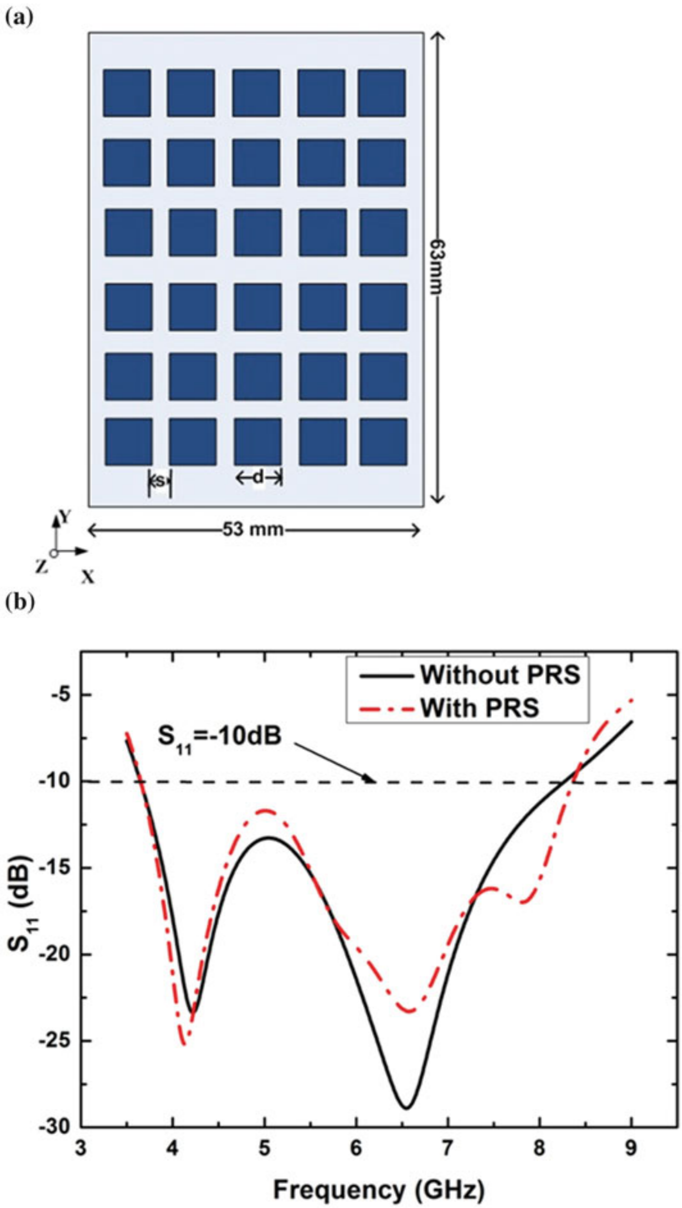


Fig. 2 **a** Design of PRS (not as per scale), **b** effect of embedding PRS on return loss of the antenna

Table 1 Comparison with earlier reported gain enhancement techniques with our proposed technique

References no.	Size of antenna (mm^2)	Frequency range	Gain (dB)	Technique used	Gain bandwidth product (dB-GHz)
[16]	50×50	530 MHz	4.62	Laminated conductor layers	2.44
[17]	341×341	860–960 MHz	9.7 dBi	Parasitic radiation patch	0.97
[18]	150×300	902–928 MHz	8.7 dBi	Metallic plane	0.22
[19]	11 layers of 60×60	8–9 GHz	7–16.4 dBi	Microwave lens	16.4
[20]	Surface area = 2100	1.54 GHz	5.5	High impedance surface	8.47
Present work	Size of PRS = 63×53	4.71 GHz (3.65–8.36 GHz)	Average 5.25 dB	PRS	24.72

that waves emanating from the PRS are in phase in normal direction, which gives more directive radiation and hence the gain enhancement can be achieved [15]. As per result shown in Fig. 3a, it is concluded that 1–2.7 dB gain enhancement achieved after embedding PRS. Between 4 and 5 GHz frequency band, the reflection coefficients are not in phase and different wave emanating from the PRS have more phase mismatch, which results less enhancement of the gain, (i.e., around 1 dB) and less radiation efficiency as shown in Fig. 3b. As frequency increases more than 5 GHz, there is a decrement in phase differences between different waves emanating from the PRS and hence more enhancement in gain as well as radiation efficiency achieved.

Table 1 shows comparison of different gain enhancement techniques of planar antennas which were reported earlier with our proposed technique of PRS. From earlier reported works, the maximum bandwidth and maximum gain using different gain enhancement techniques have been considered in this table. Our proposed UWB monopole antenna exhibits a bandwidth of 3.65–8.36 GHz with a distributed gain of 3.0–7.5 dB. In the comparative Table 1, for our work, average gain of 5.25 dB has been considered. Using our proposed gain enhancement technique (PRS), a maximum gain bandwidth of product of 24.72 dB-GHz obtained, which is better as compared to reported papers given Table 1.

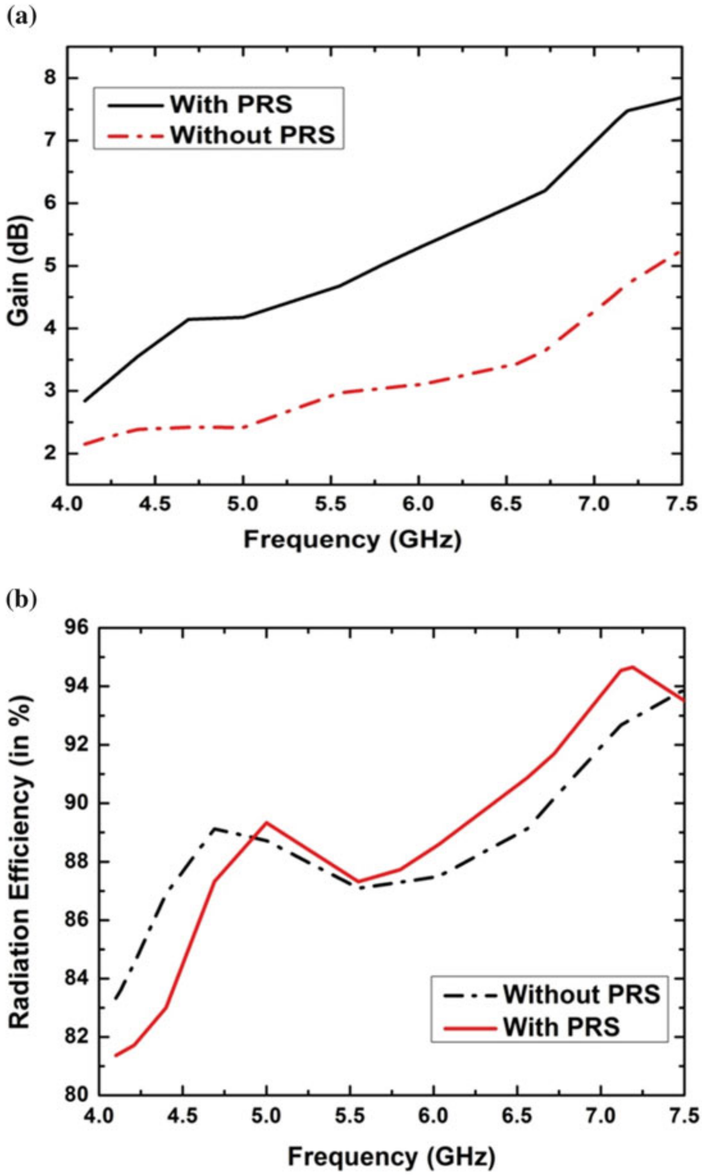


Fig. 3 Design effect of embedding PRS on antenna performance parameters. **a** Gain, **b** radiation efficiency

4 Conclusion

A microstrip line feed UWB monopole antenna embedded with PRS has been proposed for enhancement of the gain of the antenna. By optimization of the resonance distance between radiating element and PRS, the shape of copper cells, the space between copper cells and their dimensions, the gain of the antenna increases to 1–2.7 dB in UWB range. The proposed antenna gives the solution of the poor gain problem in microstrip line feed UWB monopole antenna with the cost of increment of size of antenna vertically. For further enhancement of the gain, one more layer of PRS with suitable size and numbers parasitic patches can be considered. This may be the future scope of this work.

Acknowledgements The author would like to thank the Gujarat Council on Science and Technology, Gandhinagar, Gujarat, India for funding under minor research project program. Authors are also thankful to Charutar Vidyamandal and the management of ADIT, Gujarat, India for their motivation and support.

References

1. Allen, B., Dohler, M., Okon, E.E., Malik, W.Q., Brown, A.K., Edwards, D.J.: Ultra-Wideband Antennas and Propagation For Communications, Radar and Imaging. Wiley, England (2006)
2. Iyer, B., Garg, M., Pathak, N., Ghosh, D.: Concurrent dual-band RF system for human respiration rate and heartbeat detection. In: IEEE Conference on Information and Communication Technologies (ICT), pp. 749–753 (2013)
3. Iyer, B., Pathak, N., Ghosh, D.: Dual-input dual-output RF sensor for indoor human occupancy and position monitoring. IEEE Sens. J. **15**(7), 3959–3966 (2015)
4. Li, Y.T., Yang, X.L., Li, Z.B., Wang, L., Yang, H.C.: A unidirectional cylindrical conformal monopole antenna designed for impulse radar system. IEEE Antenn. Wirel. Propag. Lett. **10**, 1397–1400 (2011)
5. Liang, J., Chiau, C.C., Chen, X., Parini, C.G.: Study of a printed circular disc monopole antenna for UWB systems. IEEE Trans. Antenn. Propag. **53**(11), 3500–3504 (2005)
6. Dikmen, C.M., Cimen, S., Cakir, G.: Planar octagonal-shaped UWB antenna with reduced radar cross section. IEEE Trans. Antenn. Propag. **62**(6), 2946–2953 (2014)
7. Chen, D., Cheng, C.H.: A novel compact ultra-wideband wide slot antenna with via holes. Progress Electromag. Res. **94**, 343–349 (2009)
8. Ranadkaew, T., Rakluea, P.: A compact moon shaped super wideband thin-film antenna. In: Proceedings of 13th International Conference on Electrical Engineering/Electronics, Computer, Tele-communications and Information Technology (ECTI-CON), Chiang Mai, pp. 1–4 (2016)
9. Lui, W.J., Cheng, C.H., Zhu, H.B.: Experimental investigation on novel tapered microstrip slot antenna for ultra-wideband applications. IET Microwave, Antennas & Propagation, vol. 1, no. 2, pp. 480–487 (Apr 2007)
10. Xu, H.Y., Zhang, H., Lu, K., Zeng, X.F.: A holly-leaf-shaped monopole antenna with low RCS for UWB application. Progress Electromagn. Res. **117**, 35–50 (2011)
11. Lin, S., Cai, R.N., Huang, G.L., Wang, J.X.: A miniature UWB semicircle mono pole printed antenna. Progress Electromagn. Res. Lett. **23**, 157–163 (2011)
12. Lin, C.C., Chuang, H.R.: A 3–12 GHz UWB planar triangular monopole antenna with ridged ground-plane. Progress Electromag. Res., vol. 83, pp. 191–198 (2008)

13. Zaker, R., Ghobadi, C., Nourinia, J.: A modified microstrip-FED two-step tapered monopole antenna for UWB and WLAN applications. Progress Electromagn. Res., vol. 77, pp. 137–148 (2007)
14. CST Microwave Studio, Version 17
15. Vaidya, A.R., Gupta, R.K., Mishra, S.K., Mukherjee, J.: High gain low side lobe level fabry perot cavity antenna with feed patch array. Progress Electromagn. Res. C **28**, 223–238 (2012)
16. Ghosh, A., Das, S.: Gain enhancement of slot antenna using laminated conductor layers. In: International Conference on Devices, Circuits and Communications (ICDCCom), Ranchi, India, pp. 1–4 (2014)
17. Zhao, X., Huang, Y., Li, J., Zhang, Q., Wen, G.: Wideband high gain circularly polarized UHF RFID reader microstrip antenna and array. Int. J. Electron. Commun. **77**, 76–81 (2017)
18. Lin, Y.F., Chang, M.J., Chen, H.M., Lai, B.Y.: Gain enhancement of ground radiation antenna for RFID tag mounted on metallic plane. IEEE Trans. Antenn. Propag. **64**, 1193–1200 (2016)
19. Cheung, S.W., Li, Q.L., Wu, D., Yuk, T.I.: Microwave lens using multi-layer substrates for antenna gain enhancement. In: 10th European Conference on Antennas and Propagation (EuCAP), Davos, pp. 1–4 (2016)
20. Amiri, M.A., Balanis, C.A., Birtcher, C.R.: Gain and bandwidth enhancement of a spiral antenna using a circularly symmetric HIS. IEEE Antenn. Wirel. Propag. Lett. **16**, 1080–1083 (2017)

Reconfigurable Inset-Fed Patch Antenna Design Using DGS for Human Vital Sign Detection Application

Brijesh Iyer, Mahesh P. Abegaonkar and S. K. Koul

Abstract The manuscript reports a simple reconfigurable inset-fed Microstrip patch antenna using defected ground structure (DGS). The DGS serves two folds in the proposed design; first it eliminates the higher order resonance and second, supports the resonance at other band with the incorporation of switching diode in it. Initially, the patch antenna has been designed to operate at 5.6 GHz. With the incorporation of the switch, a resonance at 3.36 GHz was obtained. The measured prototype shows a return loss (S11) better than −12 dB in each case with a gain of 4.2 dBi and 2.2 dBi at 'OFF' and 'ON' condition, respectively. The proposed antenna is specially designed to cater the needs of portable human non-invasive vital sign detection (NIVSD) system for medical and defense applications. Low cost, high directivity and light weight are desired for antenna used in these applications.

Keywords Antenna · Human vital sign · Inset-fed · Reconfigurable · RF sensor

1 Introduction

From its first report in 1970s, radio frequency (RF) based non-invasive detection of human vital signs has drawn a huge attention of academicians and researchers owing to its use in various day to day activities, healthcare applications and military and law enforcement application [1, 2]. Multiband architecture emerged as an alternative technology to improve the performance of existing single band RF systems for non-invasive vital sign detection (NIVSD) [3]. Selection of a specific operational

B. Iyer
B. A. Technological University, Lonere 402103, India
e-mail: brijeshiyer@dbatu.ac.in

M. P. Abegaonkar (✉) · S. K. Koul
CARE, Indian Institute of Technology Delhi, New Delhi 110016, India
e-mail: mpjosh@care.iitd.ac.in

S. K. Koul
e-mail: skkoul@care.iitd.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_9

Table 1 State-of-the-art NIVSD centric antenna design

Contribution	Directivity	No. of patch element	Dimension	Reconfigurability
Park et al. [4]	Broadband	Single	50 × 30 mm	NO
Tang et at. [5]	Broadband	2 × 4 array	100 × 96 mm	NO
Iyer et al. [6]	Directive	1 × 2 array	78 × 97 mm	NO
Present Work	Directive	Single	60 × 60 mm	YES

frequency and the reduction in higher order harmonics are the key of successful deployment such systems as a sensor. Out of various front-end elements, antenna is a crucial element as it acts as a transceiver element.

In the recent past, development of dualband antenna is initiated, in large extent, for wireless applications with bidirectional radiation pattern. A non-contact vital sign detection centric approach was proposed in [4–6]. However, these activities ended with a broadband performance or a bulky size using array approach. Table 1 shows the state-of-the-art NIVSD centric antenna design approach.

The frequency reconfigurable antenna offers advantages such as compact size, similar radiation pattern and gain as compared to the multiband antennas [7–11]. The frequency reconfigurability may be achieved by using electromechanical switches like RF MEMS, PIN diode, varactor diode and FETs [12–15]. However, a PIN diode has the advantage of low cost and high power handling capacity with low driving voltage.

In the present work, a simple inset-fed reconfigurable microstrip patch antenna is proposed. The antenna is characterized with the use of an inverted U-shaped slot just below the feed line in the ground plane and the incorporation of PIN diode as a switch. The inverted U-shaped slot helps to eliminate the unwanted higher order harmonics and the PIN diode aids the reconfigurability of the frequency.

Section 2 of the paper describes the geometry of the antenna, while the fabrication of the proposed antenna prototype, its surface current distribution, return loss and radiation pattern has been given in Sect. 3. The paper is concluded with its likely benefit and future expansion.

2 Geometry of the Proposed Antenna

Figure 1 shows the geometry of the antenna along with DGS. The rectangular patch is designed, to operate at 5.6 GHz, using the empirical relations given in [11]. Commercially available electromagnetic simulation tool CST microwave studio has been used in this study to obtain optimized characteristics of the antenna.

Roger RT5880 substrate of dimension (L × W) 60 × 60 mm with dielectric constant $(\varepsilon_r) = 2.2$, substrate height $(h) = 0.762$ mm with a thickness $(t) = 0.016$ mm is used for the proposed design. The patch dimensions are: Length of patch (PL) =

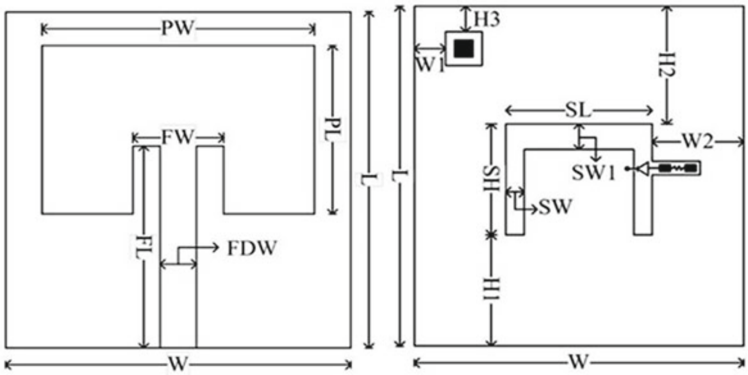


Fig. 1 Geometry of the proposed antenna: **a** Patch side **b** Ground side

19.24 mm, width of the patch (PW) = 22 mm, feed length (FL) = 5.7 mm, feed width (FDW) = 2.38 mm and inset-feed width (FW) = 3.57 mm. The ground side is characterized by an inverted U shaped DGS. The DGS dimensions are: SL = 7 mm, SH = 8 mm, SW = 0.4 mm and SW1 = 0.5 mm. The geometric dimensions of the ground side are: H1 = 21 mm, H2 = 31 mm, H3 = 10 mm, W1 = 20 mm and W2 = 26 mm. A pin diode is incorporated as a switch in right arm of DGS. Pads of 2 × 2 mm dimensions are used for providing the supply to the diode.

3 Results and Discussions

According to the initial design consideration, the antenna operates at 5.6 GHz band. However, it has been observed that along the fundamental design frequency, its harmonics are also present. Such characteristic is undesirable. Figure 2 shows the simulated return loss characteristics of the proposed antenna.

To overcome this issue, an inverted U-shaped slot is created in the ground plane. This structure may act as a DGS. The DGS is placed exactly below the feed line completely reduced the unwanted harmonics from the operation. Further, the antenna is converted into a frequency reconfigurable structure by using a PIN diode in the DGS. The operation of the proposed antenna has been verified in two different operational conditions i.e. switch is ON and switch is OFF. Figure 3 shows the fabricated prototype of the proposed antenna.

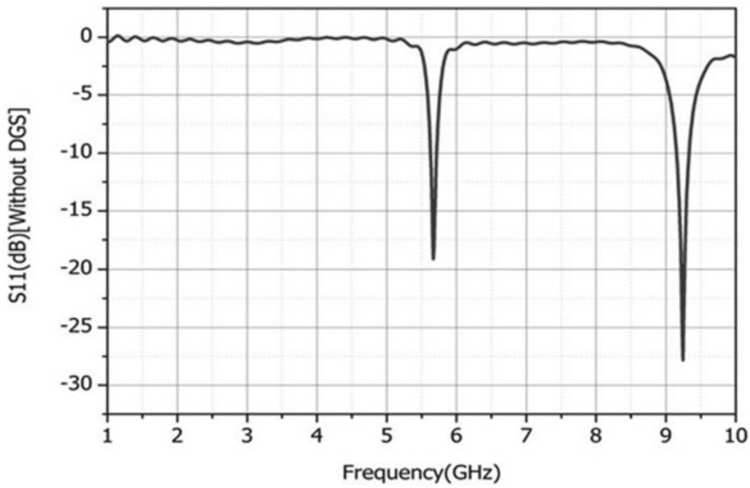


Fig. 2 Return loss characteristics of the antenna without DGS

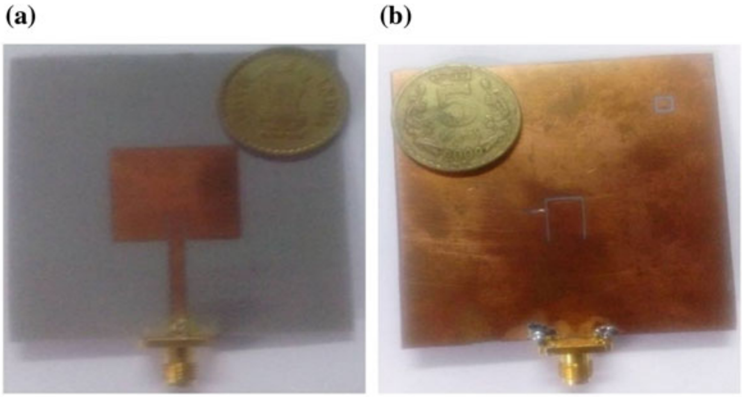


Fig. 3 Fabricated prototype of the proposed antenna: **a** Patch side **b** Ground side

3.1 When Switch Is in 'OFF' Condition

In this state, the antenna operates at 5.6 GHz band as per the basic design parameters. The operation has been confirmed by the close agreement between simulated and measured return loss and radiation pattern of the proposed antenna. Figure 4 shows the simulated return loss characteristics and surface current distribution of the proposed antenna. The radiation pattern is depicted in Fig. 5.

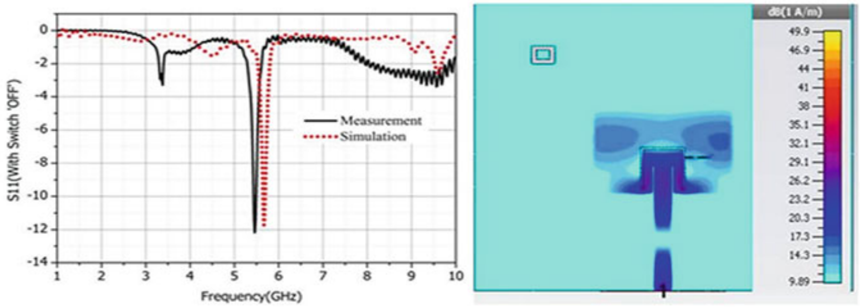


Fig. 4 Characterization at switch OFF condition: **a** Return loss **b** Surface current distribution

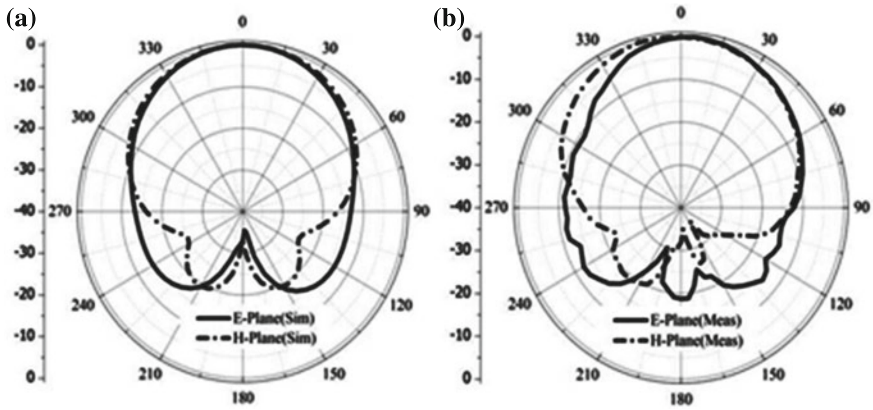


Fig. 5 Radiation pattern at switch OFF condition: **a** Simulation **b** Measurement

3.2 When Switch Is in 'ON' Condition

In this state, the antenna operates at 3.47 GHz band due to the switching operation. It is evidenced from the surface current distribution that the proposed band exists due combined effect of DGS and the switch. From Fig. 4b it is very clear that the similar effect was absent when the switch was in the OFF state.

The operation has been confirmed by the close agreement between simulated and measured return loss and radiation pattern of the proposed antenna. Figure 6 shows the simulated return loss characteristics and surface current distribution of the proposed antenna. The radiation pattern is depicted in Fig. 7.

The *Friis* transmission equation is employed to estimate the realizable gain of the proposed patch antenna. Table 2 provides the gain characteristics of the proposed antenna. A harmonic free operation has been evidenced from the antenna characterization at both conditions. This is highly desirable in human vital sign detection applications as the vital signs are very tiny in nature and the harmonics may suppress

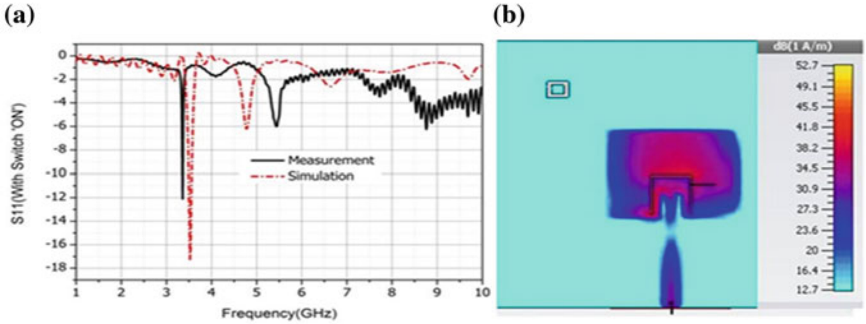


Fig. 6 Characterization at switch ON condition: **a** Simulation **b** Measurement

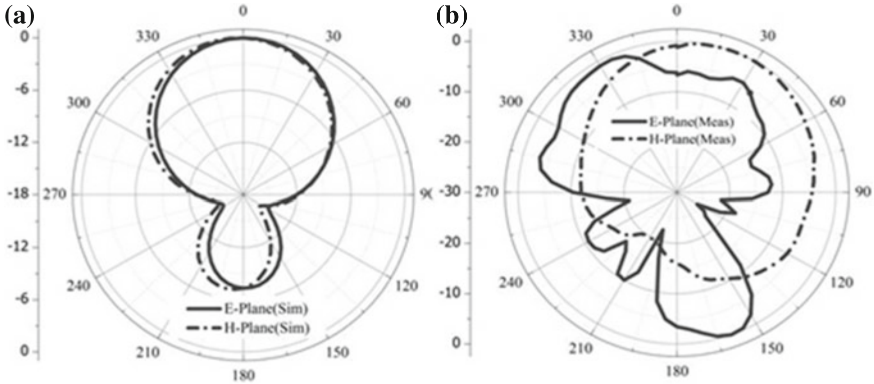


Fig. 7 Radiation pattern at switch ON condition: **a** Simulation **b** Measurement

Table 2 Gain characterization

Switch state	Frequency (GHz)	Gain (dBi)	
		Simulation	Measurement
OFF	5.6	7.7	4.2
ON	3.36	7	2.2

them. Further, more the individual band radiation pattern is highly directive with considerable gain. All these features are obtained with a very small antenna dimensions. Hence, the proposed antenna structure may be useful in the portable human NIVSD RF sensors.

The proposed antenna is very compact, low cost and easily reproducible. It does not suffer from the higher order harmonics. Park et al. [4] had reported a compact antenna structure. However, it is omnidirective radiation pattern and static in nature in terms of operative frequency. The approach in [5] suffers from the drawback of large size with omnidirective radiation pattern. The approach reported in [6] is having directive nature of radiation pattern with large size.

For non-invasive human VSD sensors, the antenna must be highly directive, low cost and compact in nature. With the same antenna prototype two different operational bands may be achieved. Hence, the proposed antenna prototype may be useful for non-invasive human vital sign detection RF sensors. In future, the proposed antenna structure may be modified for continuous in-band tuning. This may help to select an optimal frequency for the detection of tiny human vital signs for a particular application.

4 Conclusions

An inverted U slot DGS inspired frequency reconfigurable inset-fed microstrip patch antenna is designed. The resonance at ON condition is sensitive to the DGS and the switching operation. This design may be easily incorporated with microwave circuits for practical application where a directive nature of radiation is required with compact size. In future, the proposed antenna structure may be modified to achieve the in-band tuning to select an optimal frequency for a specific NIVSD application.

Acknowledgements The authors are thankful to Indian National Academy of Engineers (INAE), India for sponsoring this work in the year 2015–2016.

References

1. Li, C., Lin, J.: Random body movement cancellation in Doppler radar vital sign detection. IEEE Trans. Micro. Theory Tech. **56**(12), 3143–3152 (2008)
2. Iyer, B., Pathak, N., Ghosh, D.: Multiband concurrent RF system for human vital sign detection during natural calamities. In: IEEE International Microwave & Radio Frequency Conference (IMaRC-13), New Delhi, India, pp. 1–4 (2013)
3. Iyer, B., Pathak, N., Ghosh, D.: Dual-Input dual-output RF sensor for indoor human occupancy and position monitoring. IEEE Sens. J. **15**(7), 3959–3966 (2015)
4. Park, Z., Li, C., Lin, J.: A broadband microstrip antenna with improved gain for noncontact vital sign radar detection. IEEE Ant. Wirel. Prop. Lett. **8**, 939–942 (2009)
5. Tang, T., Chuang, Y., Lin, K.: A narrow beamwidth array antenna design for indoor non-contact vital sign sensor. In: Antennas and Propagation Society International Symposium (APSURSI-12), Chicago, IL, pp. 1–2 (2012)
6. Iyer, B., Pathak, N., Ghosh, D.: Concurrent dualband patch antenna array for non-invasive human vital sign detection application. In: Asia Pacific Conference on Electromagnetics (APACE-14), Johor-bahru, Malaysia, pp. 150–153 (2014)
7. Sharbati, V., Rezaei, P., Fakharian, M., Beiranvand, E.: A switchable band-notched UWB antenna for cognitive radio applications. IETE J. Res. **61**(4), 423–428 (2015)
8. Qin, P., Weily, A., Guo, Y., Bird, T., Liang, C.: Frequency reconfigurable Quasi-Yagi folded dipole antenna. IEEE Trans. Antenn. Propag. **58**(8), 2742–2747 (2010)
9. Hum, S., Xiong, H.: Analysis and design of a differentially-fed frequency agile microstrip patch antenna. IEEE Trans. Antenn. Propag. **58**(10), 3122–3130 (2010)

10. Prajapati, P.: Multilayered circularly polarized microstrip antenna integrated with defected ground structure for wide impedance and axial ratio bandwidth. J. Electromag. Waves Appl. **30**(17), 2256–2267 (2016)
11. Cetiner, B., Crusats, G., Jofre, L., Biyikli, N.: RF MEMS integrated frequency reconfigurable annular slot antenna. IEEE Trans. Antenn. Propag. **58**(3), 626–632 (2010)
12. Nikolaou, R., Lugo, C., Carrasquillo, I., Ponchak, G.: Pattern and frequency reconfigurable annular slot antenna using PIN diodes. IEEE Trans. Antenn. Propag. **54**(2), 439–448 (2006)
13. Shynu, S., Augustin, G., Mohanan, P., Vasudevan, K.: Design of compact reconfigurable dual frequency microstrip antenna using varactor diodes. Prog. Electromag. Res. **60**, 197–205 (2006)
14. Aboufoul, T., Alomainy, A., Parini, C.: Reconfigurable UWB monopole antenna for cognitive radio applications using GaAs FET switches. IEEE Antenn. Wirel. Propag. Lett. **11**, 392–394 (2012)
15. Balanis, C.: Antenna Theory, Analysis and Design. Wiley, Inc., New York, Publisher (1997)

Asymmetric Double U-Slot Multi-frequency Antenna for WLAN/5G Communication

Sraddhanjali Mohapatra, Debaprasad Barad and Subhrakanta Behera

Abstract A compact microstrip planar antenna with asymmetrical slot width on single dielectric layer has been investigated for multi-frequency operation. The antenna is especially designed for 5G communication and WLAN communication. The geometry of the proposed antenna comprises of a single dielectric layer with a single radiating element, in which the rectangular patch is introduced with dual U-Slot to achieve multiple resonances. The antenna is excited by a transmission line feeding, presents on the same layer. The antenna first resonates at 4.2 GHz, later resonates at 5.2 and 5.8 GHz. Proposed patch configuration shows improved bandwidth of 180, 350, and 250 MHz with the offset transmission line feed and asymmetric slot. Both symmetric and asymmetric behavior of dual U slot has been studied. The antenna was fabricated using a single FR4 substrate of dimension $16 \times 20 \times 1.56\ \text{mm}^3$.

Keywords Single dielectric layer · Multiple frequency · Single port
Transmission line feed · 5G · WLAN · Wi-MAX

1 Introduction

In this era of wireless communication system, the electronic devices usually need antennas to operate for a variety of signals. With the expansion of information industry has emerged the miniaturization of the electronics products, so that an antenna with multi-frequency application capacity, compact size, high efficiency and low cost has developed into an essential basic for the unification of wireless communication system. An antenna with multi-band operation can facilitate the multi-tasking communication system. The Microstrip patch antenna is popular for its excellent characteristics like low cost, light weight, easy fabrication and facility to print on the same board as transceiver or receiver. For achieving the same, the interest goes towards the development of more and more RF bands into a single chip [1–3].

S. Mohapatra · D. Barad (✉) · S. Behera
Kalinga Institute of Industrial Technology (Deemed to be University), Bhubaneswar 751024, India
e-mail: deba7482@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_10

In [4], a compact U-slot antenna is proposed WLAN/Wi-max communication with dimension 10×26 mm^2. A metamaterial inspired triple band antenna [5] is proposed for reconfigurable multi-band application. The said contribution is somewhat complex in configuration. The work mainly contributes a compact microstrip planar antenna module applicable for multi-functional wireless communication. Particularly the antenna is usable in mobile communication such as under 6 GHz *5G* communications and also in WLAN/Wi-Max Communication [6, 7].

In this work, a compact microstrip-fed antenna using a single radiating element with dual U slot for tri-band operation is presented. Integration of multiple techniques on antenna patch reduces the physical size of the antenna and enhances the antenna performance. Parametric study of the antenna with symmetrical and asymmetrical slot width has been reported for achieving better resonance characteristics. The show resonances at 4.2, 5.2, and 5.8 GHz. Also exhibits good gain response of maximum 4 dBi.

2 Antenna Design

The proposed tri-band antenna is designed using a single FR4 dielectric substrate having a dielectric constant (ε_r) 4.4, thickness of (h_1) 1.56 mm, loss tangent ($\tan \delta$) of 0.02. This single layer architecture uses a dual side laminated substrate in which the antenna geometry is implanted on top surface and the ground was implanted on the bottom surface.

2.1 Antenna Configuration

The conventional antenna structure comprises of a rectangular radiator patch of dimension 16×20 mm^2 with a microstrip-fed transmission line feed of dimension 9.5×3 mm^2. In this discussed architecture, both the antenna radiator and feed line is placed on the same layer and a conductive plane is considered as ground plane, which is placed below to the antenna patch as demonstrated in Fig. 1. In which, Fig. 1a depicts the cross-section view and the Fig. 1b depicts the geometry of the proposed antenna. The fullwave simulation of the designed antenna has been carried out using Method of Moment (MoM) EM solver.

In the beginning a regular rectangular planar structure is considered as a conventional antenna patch. Wherein, a simple transmission line is considered for exciting the radiator patch. The designed patch is able to contribute single RF band. Later, the conventional structure is optimized with symmetric and asymmetric U-slot to acquire the desired multi-resonance characteristics. The physical dimension of the proposed antenna is given in Table 1.

The physical dimensions of the antenna are determined from following equations [8].

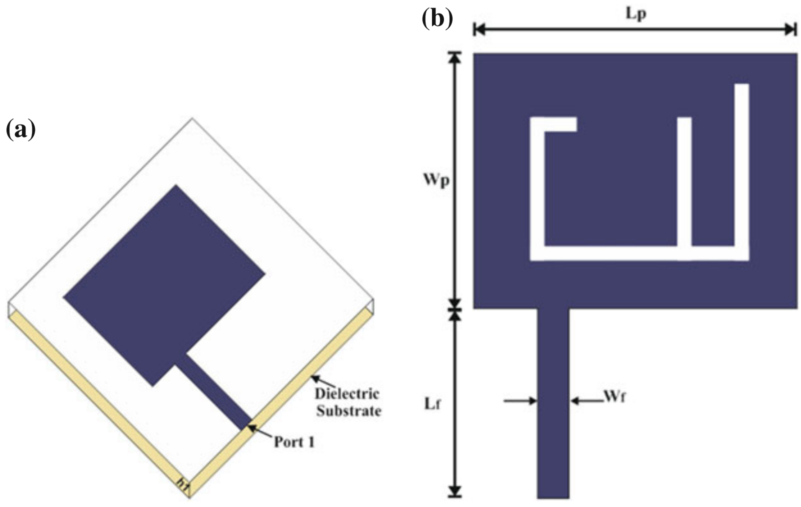


Fig. 1 Configuration of single-layer microstrip-fed antenna for multiple frequency operation. **a** Cross view. **b** Top view

Table 1 Physical dimensions of the proposed antenna

Parameters	Descriptions	Values
ε_r	Dielectric constant	4.4
h_1	Thickness of substrate	1.56 mm
$\tan\delta$	Loss tangent	0.0018
Lp	Length of the patch	20 mm
Wp	Width of the patch	16 mm
Lf	Length of the feed	9.5 mm
Wf	Width of the feed	3 mm
U_L	Length of the U-slot	8 mm
Uw	Width of the U-slot	1 mm

$$W = \frac{1}{2f_r\sqrt{\mu_0\epsilon_0}} \times \sqrt{\frac{2}{\epsilon_r + 1}} \tag{1}$$

$$L = \frac{1}{2f_r\sqrt{\epsilon_{eff}}\sqrt{\mu_0\epsilon_0}} - 2\Delta L \tag{2}$$

$$\epsilon_{eff} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2\sqrt{1 + 12\frac{h}{W}}} \tag{3}$$

The optimized dimensions of the antenna are acquired during the full wave simulation and from [9].

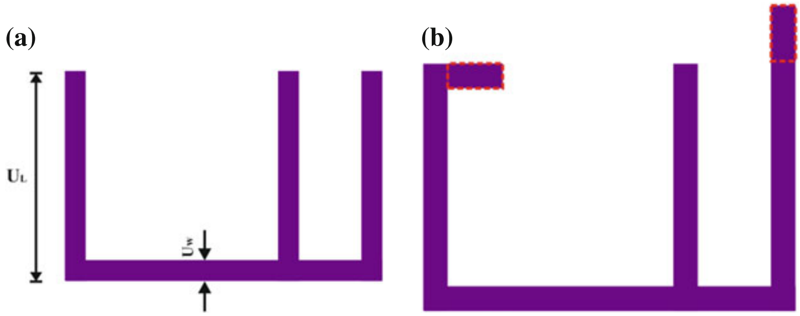


Fig. 2 Configuration of dual U-slot for multi-frequency operation. **a** Dual U-slot. **b** Optimized dual U-slot

The configuration of the slot architecture comprises of two U-slots of length U_L and width Uw. In the beginning, the antenna characteristics are observed with this symmetrical dual U-slot as shown in Fig. 2a. Later, the existed dual U-slot is optimized with stub loading. The newly loaded stub is placed in the suitable top portion and marked with a red dash line as shown in Fig. 2b. This stub of dimension 2.5×1 mm^2 makes the U-slot asymmetric.

During the first iteration, a symmetric U-Slot is introduced on the antenna patch. Even if with this optimization the resonance characteristics is improved, but the required characteristic is not achieved. Again, another symmetric U-slot is added alongside the first U-slot. This combines development makes a dual U-slot configuration which is able to provide multi-resonance characteristics. But the antenna still needs improved impedance matching and also enhancement of the gain characteristic is desired. The discussed dual U-slot architecture is optimized periodically to improve the antenna characteristics. The rigorous analysis has been concluded that the asymmetric configuration of the dual U-slot yields improved impedance characteristic and the gain response of the proposed antenna is enhanced.

3 Result and Discussion

The proposed antenna with asymmetrical double U-slot has been investigated rigorously; the corresponding results are discussed in this section. Figure 3 depicts the resonance characteristics. From this it is clearly demonstrates that the proposed antenna exhibits, multi-resonance characteristics. The resonance characteristics comprise of two graphs; the dashed graph represents the S_{11} parameter of conventional rectangular patch and the solid graph represents the S_{11} parameter of the proposed antenna patch which is well below -10 dB reference line. The proposed antenna shows improved operational bandwidth of 4.1–4.28 GHz at first resonance, 4.95–5.3 GHz at second resonance, and 5.68–5.88 GHz at third resonance However, the antenna

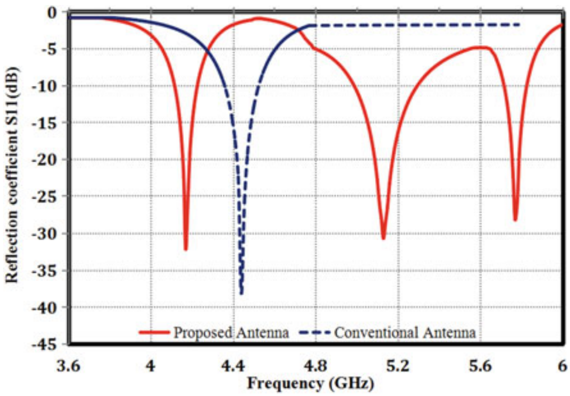


Fig. 3 Resonance characteristics of the proposed single-layer antenna

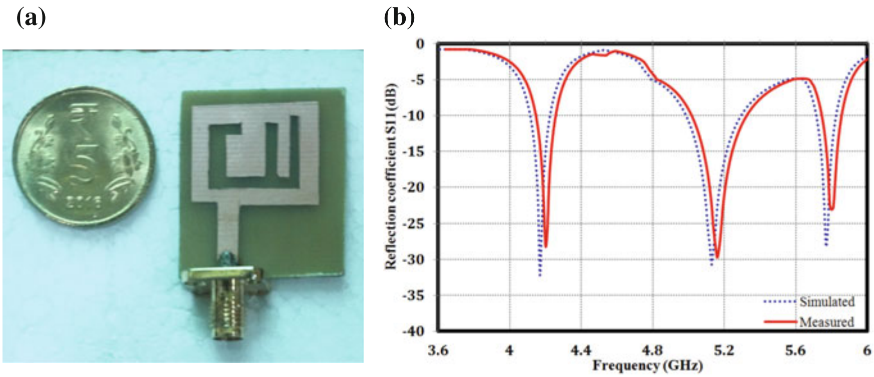


Fig. 4 Experimental study of the proposed multi-band antenna. **a** Fabricated prototype. **b** Measured S_{11} parameter

shows improved impedance bandwidth of nearly 350 MHz. It has been observed that the antenna yields good gain response at each resonance as demonstrated in Fig. 4. The proposed antenna shows gain response up to 4 dBi.

The proposed antenna geometry is fabricated using a single FR4 substrate as discussed in Sect. 2. The proposed geometry is implemented on one side of the substrate following the photolithography process, wherein the other side is considered as the ground plane. A Sub miniaturized Version A (SMA) connector is mounted on the microstrip line feed at the extreme bottom of the same layer. The final fabricated prototype of the proposed antenna is depicted in Fig. 4a. It has been clearly demonstrated that the prototype is comparable to a coin.

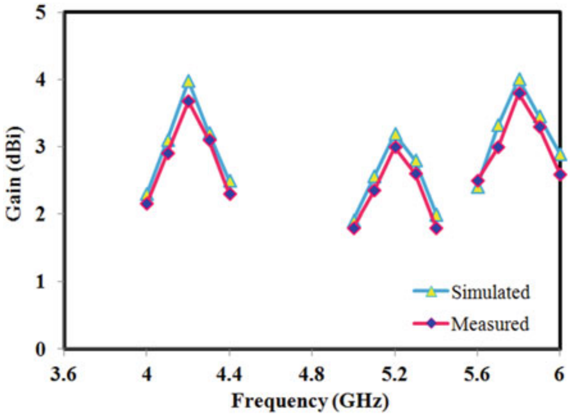


Fig. 5 Gain response of the proposed antenna

The antenna resonance characteristic is observed from the Agilent Vector Network Analyzer (VNA). From Fig. 4b, it has been clearly observed that the measured result is well agreed with each other. The antenna first resonates at 4.2 GHz, later it resonates at 5.2 and 5.8 GHz. The antenna shows an improved band of operation with the asymmetric slot in the center of the patch.

The asymmetric slot structure contributes more current path, as a result maximum edge excitation occurs. This extra edge radiation improves the output characteristics such as enhanced gain response at each resonance. It has been observed that the antenna yields good gain response at each resonance as demonstrated in Fig. 5. The proposed antenna shows gain response up to 4 dBi. Also, it has been realized the measure gain response has a good agreement with the simulated result.

The radiation pattern of the proposed antenna is shown in Fig. 6. The radiation characteristics at first resonance are shown in Fig. 6a, the corresponding characteristics at second and third resonances are shown in Fig. 6b and Fig. 6c respectively.

However, owning all the notable characteristic the single-layer, single-feed planer antenna with asymmetrical double U-slot is proposed for multi-frequency operation with improved output characteristics.

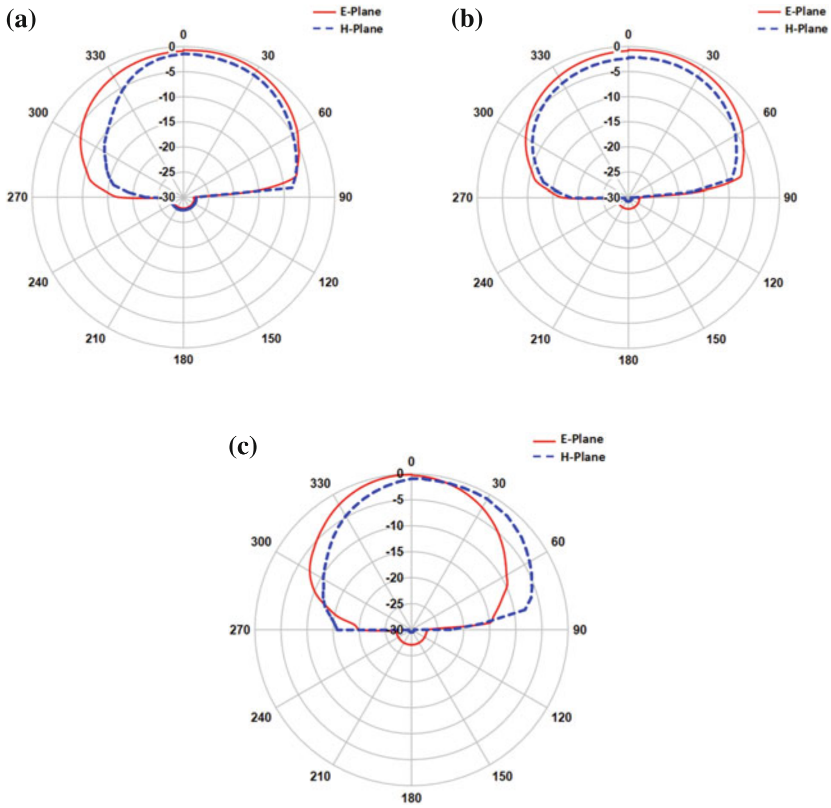


Fig. 6 Radiation pattern at each resonance. **a** Radiation pattern at 1st resonance. **b** Radiation pattern at 2nd resonance. **c** Radiation pattern at 3rd resonance

4 Conclusion

A Compact Linearly polarized microstrip antenna for multi-frequency WLAN and 5G application has been presented in this study. The asymmetric behavior of dual U slot at radiator patch is carefully analyzed for achieving multiple resonance characteristics. The microstrip-fed transmission line fed enhances the impedance profile of the single layer antenna, wherein the impedance bandwidth 180 MHz at 4.2 GHz, 350 MHz at 5.2 GHz, and 200 MHz at 5.8 GHz is realized. The proposed antenna with slotted architecture yields good gain response of 4dBi, 3.4dBi, and 4.2dBi respectively. However, owning all these characteristics the compact antenna is proposed for multi-functional 5G communication and Wi-MAX/WLAN communication system along with C-Band wireless application.

References

1. Park, D.-H., Kwak, Y.-S.: Design multi-band microstrip patch antenna for wireless terminals. IEEE Futur. Gener. Commun. Netw. Lett. **2**, 439–441 (2007)
2. Balanis, A.: Antenna Theory Analysis and Design, 3rd edn. Wiley, Inc Publication, UK (2003)
3. Costantine, J., Kabalan, K.Y., Ei-Hajj, A., Rammal, M.: New multi-band microstrip antenna design for wireless communications. IEEE Antenn. Propag. Mag. **49**(6) (2007)
4. Kunwar, A., Gautam, A.K., Rambabu, K.: Design of a compact U-shaped slot triple band antenna for WLAN/Wi-Max applications. Int. J. Electron Commun. (AEU) **71**, 82–88 (2017)
5. Rajeshkumar, V., Raghavan, S.: A compact metamaterial inspired triple band antenna for reconfigurable WLAN/WiMAX application. Int. J. Electron Commun. (AEU) **69**, 274–280 (2017)
6. Optimizing 4 GHz mission critical networks. Microw. J. Radio Waves, Billerica, MA (2009)
7. Yurduseven, O., Smith, D., Pearsall, N., Forbes, I.: A solar cell stacked slot-loaded suspended microstrip patch antenna with multi-band resonance characteristics for WLAN and WIMAX systems. Progress Electromagn. Res. **142**, 321–332 (2013)
8. Garg, R., et al.: Microstrip Antenna Design Handbook. Artech-house, Boston, London (2001)
9. Mohapatra, S., Barad, D., Behera, S.: Realization of AeroMACS network adopting compact microwave antenna. Int. J. Microw. Opt. Technol. **12**(4), 268–274 (2017)

Performance Analysis of Optimal Versus Energy-Based Selection of Receiver Antenna for MIMO Systems

Nitin Deotale and Uttam Kolekar

Abstract The paper reports the performance analysis of antenna selection at receiver end. The optimal and energy-aware antenna selection techniques are compared for multiple input multiple output (MIMO) antenna systems. The analysis is carried out with a very simple cost function, i.e., CDF. The performance analysis is carried out with Simulink tool of MATLAB. The plots of CDF versus channel capacity and bit error rate (BER), for each case, are taken into consideration for the analysis of antenna selection. It is found that the antenna selection at the receiver side is more impactful in MIMO system.

Keywords Antenna selection · CDF · Channel capacity · MIMO · SNR

1 Introduction

The channel capacity in the communication system can be drastically improved by using multiple transceiver antennas. However, the system will become very complex, bulky, and costly due to the requirement of RF front ends for every transceiver antenna. This drawback can be overcome by using antenna selection mechanism, i.e., to use only a specific set of antenna. This approach reduces the implementation cost and circuit complexity with increased operational reliability. The antenna selection strategy may be employed at the transmitting/receiving or simultaneously at both ends. Antenna selection at transmitter side benefits a better synchronization of selection mechanism at receiver side. However, it suffers from the static nature of the operation. The antenna selection at receiver side eliminates the staticness involved in transmitter selection mode but at the cost of exhaustive search for the best fit among

N. Deotale (✉)
Priyadarshini CoE, Nagpur, India
e-mail: nitindeotale32@gmail.com

U. Kolekar
A. P. Shah Institute of Technology, Thane, Mumbai, India
e-mail: uttamkolekar@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_11

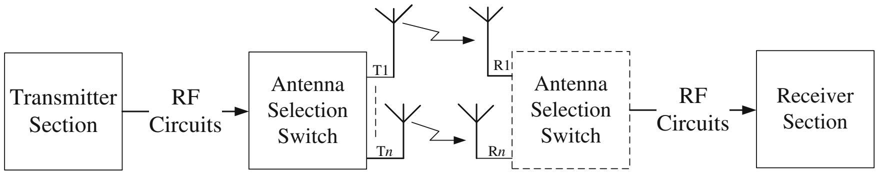


Fig. 1 MIMO system with antenna selection

all available antennae. Hence, there is a trade-off between the selection of antennas in MIMO system. The present work is motivated from this trade-off and attempts to put forward a very simple technique for antenna selection at receiver side. Multiple input multiple output (MIMO) was first reported in 1987 with computer simulations [1]. Since its first report, MIMO antenna occupied the attention of academicians and industry for research investigation due to its suitability for next-generation communication systems. Figure 1 shows the generalized diagram of a MIMO system with antenna selection.

The rest of the paper is organized as follows: Sect. 2 describes the experimental verification of the MIMO antenna selection criteria. The paper is concluded in Sect. 3 with a discussion of future research direction.

2 The Technology Overview

Many notable works have been reported various approaches for the MIMO antenna selection. Maximization of channel capacity and received SNR are the two dominant selection criteria used in MIMO antenna selection. *S. Sanayei* et al. reported a MIMO selection technique in [2]. MIMO transmit antenna selection based on channel condition was proposed by *Deotale* et al. in [3]. Adaptive MIMO antenna selection via discrete stochastic optimization was discussed in [4]. Antenna selection and power allocation techniques at transmitter over correlated channels for MIMO applications were discussed in [5]. However, very few contributions of the report describe the best policy of MIMO antenna selection and its location. Table 1 summarizes the state of the art in MIMO antenna selection.

Table 1 Brief comparison of MIMO antenna selection techniques

Contribution	Strategy	Operational side
Deotale et al. [3]	Orthogonal block coding	Transmitter
Berenguer et al. [4]	Discrete stochastic optimization	Transmitter
Zhang et al. [6]	Cross-entropy optimization	Receiver

In the present work, a comparison between optimum selection and energy-based selection strategies is carried out with an aim to predict the best location for antenna selection. MATLAB-based MIMO model is developed for the experimental verification. The cumulative distribution function (CDF) is used to verify the channel capacity.

3 The Strategies for Antenna Selection

3.1 Optimal Selection

Consider a MIMO system in which T_t = the transmitter antennas, Rr = the receiver antennas, I_t = number of selected transmitter antenna, I_r = number of selected receiver antenna. Under null antenna selection condition, $I_t = T_t$ and $I_r = Rr$. The signal received by antenna a at time t is given by $y_a(t)$ and written as

$$y_a(t) = \sqrt{\rho \sum_{i=1}^{N_t} x_i(t)h_{i,a} + n_a(t)}, \tag{1}$$

where h is the fading coefficient between a specific transmitter and receiver antenna and n is the complex Gaussian noise corresponding to the receiver antenna.

The signal received by all receiver antennae can be written as

$$y = \sqrt{\rho} x \boldsymbol{H} + \boldsymbol{n}, \tag{2}$$

where $\boldsymbol{H}$ is a *matrix* to show the simultaneous signal transmission from available antennas and $\boldsymbol{n}$ = the noise sample vector.

Under the absence of CSI at the transmitting end, the maximum channel capacity can be obtained only if the antenna power is distributed equally among the available antenna. The capacity of a MIMO system with all antennas is given as follows:

$$C = \log det\left(I_{Tr} + \frac{\rho}{Rr}\boldsymbol{H}^H \boldsymbol{H}\right). \tag{3}$$

To maximize the capacity at receiver end, the receiver selects best Rr antennas. Hence, the resulting capacity can be given as

$$C_{sel} = \arg max\, s\left(\tilde{\boldsymbol{H}}\right)\left(\log det\left[I_{Rr} + \frac{\rho}{Tr}\tilde{\boldsymbol{H}}^H \tilde{\boldsymbol{H}}\right]\right), \tag{4}$$

where $\tilde{\boldsymbol{H}}$ is calculated by omitting the unwanted antenna calculation columns from **H** and $s\left(\tilde{\boldsymbol{H}}\right)$ = Set of all such matrices.

To reduce the computational complexity for optimum antenna selection, many notable works [2–4] have been reported so far. An experimentation is carried out for

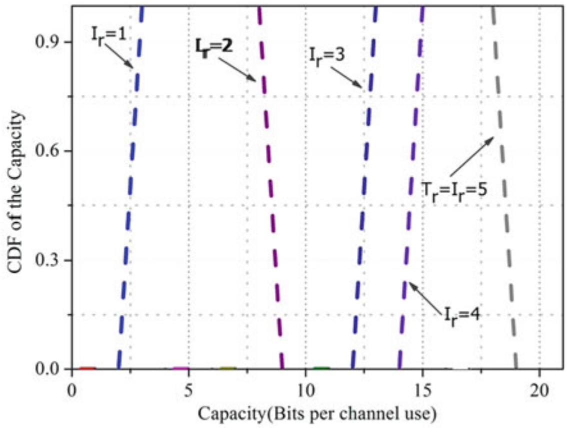


Fig. 2 CDF of the capacity for a MIMO channel at SNR = 20 dB, $T_t = 3$, $R_r = 5$, and $I_r = 1–4$

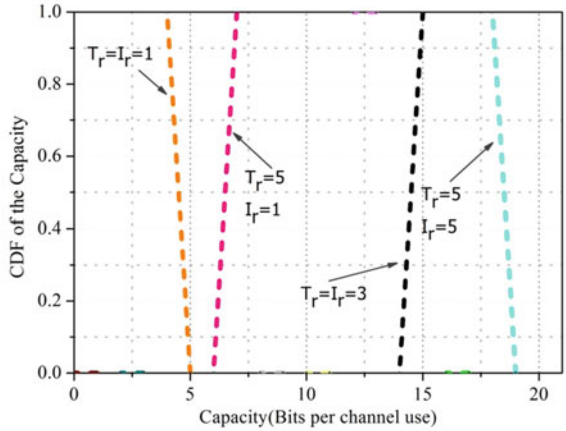


Fig. 3 CDF of a MIMO channel capacity at SNR = 20 dB, $T_r = 3$, $R_r = 1, 3, 5$, and $I_r = 1, 3$

the verification of channel capacity with $T_t = 3$, $R_r = 5$, and $I_r = 1–4$ with SNR of 20 dB with 5% outage capacity. The capacity is verified by plotting CDF. Figure 2 shows that when the selection mechanism is employed, a capacity of around 25 bits per channel use is achieved (when $R_r = I_r = 5$). A decrease to 23 bits and 21 bits per channel is observed when $I_r = 4$ and 3, respectively. This is not a considerable drop- when complexity reduction is considered. The noticeable reduction in channel capacity is observed when $Ir = 1$ and 2. This phenomenon can be credited to the fact that the independent channels are not possible simultaneously when $I_r < T_{t.}$ The observations in Fig. 2 are confirmed in Fig. 3.

Figure 3 depicts the CDF plot of the capacity for certain cases with a cap for antenna selection. In all cases, the number of transmit antennas is $T_t = 3$. Here, the capacity when $Nr = L_r = 1$ is lower to the capacity for $Nr = 5$, $Lr = 1$ by only 1.5 bits/channel use. This condition is the result of $Lr < Nt$, and thus the capacity is set by Lr. In disparity, when $Nr = Lr = 3$, the capacity is inferior in comparison with

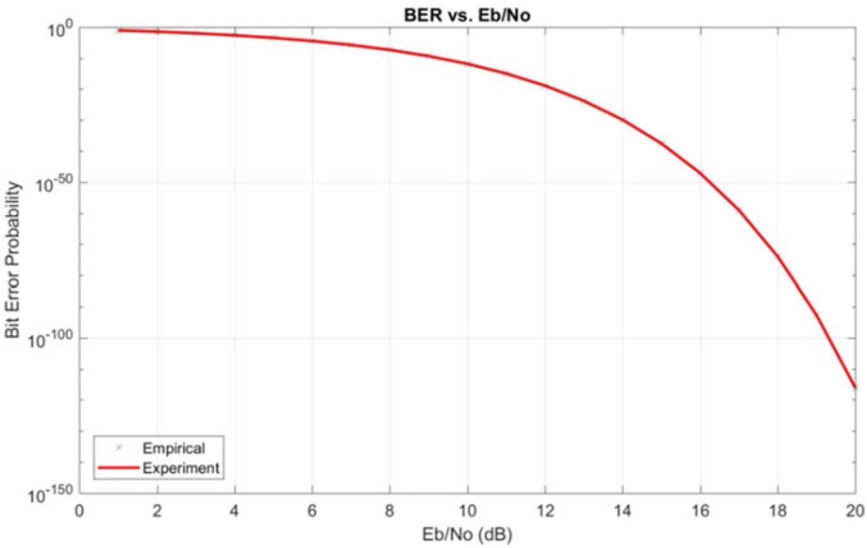


Fig. 4 BER analysis for optimum channel analysis

the condition $Nr = 5$, $Lr = 3$ by about 3 bits/channel use. Figure 4 shows BER plot of optimal selection method.

3.2 Energy-Based Selection

The received SNR can be improved by energy-based antenna selection in MIMO systems. This approach is attractive and feasible too as it is a practical one in comparison with the channel capacity maximization approach.

This strategy has been found applicable when the number of selected antennas is $Lr = 1$. For other conditions, increasing SNR, generally, does not result in the enhancement of the channel capacity. Hence, this criteria may lead to a capacity loss for all the cases, which do not satisfy $Lr = 1$. This indicates that the capacity is more susceptible to the number of antenna elements and allied phase shifts rather than the SNR.

Let us select two antennas, with highest gain, from the available Nr antennas. However, the channel gain of these antennas and the remaining antennas may be the same. As a consequence, output capacity will be less in spite of the maximization of SNR. On the other hand, if antenna pickup is initiated on the basis of maximum capacity, then the optimal selection algorithm will select one of these antennas (with high gain). The second antenna may be selected from the pool of the unselected

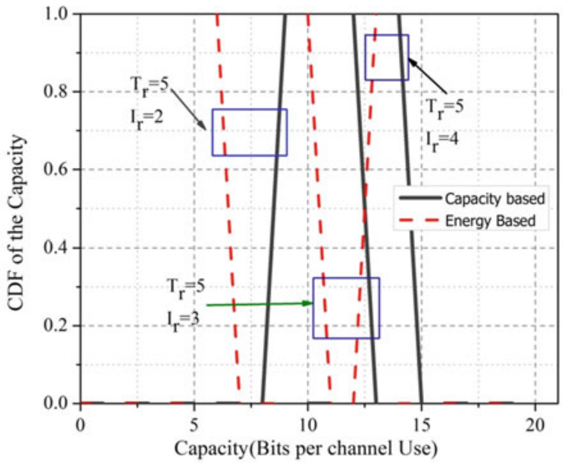


Fig. 5 Performance analysis of CDF of the capacity for a MIMO channel

Table 2 CDF plot analysis

Selection approach	Lr	Effect on channel capacity	Outage capacity
Energy based	4	Worsens by 0.5 bits/channel use	10^{-2}
	3	Loss increases to 1.8 bits/channel use	
	2	Loss reduces to 1.2 bits/channel use	

(remaining) antennas. This approach may result in the selection of a high-rank channel with improved capacity than the random selection strategy.

Figure 5 shows a CDF plot of the capacity of a $(Tr, Rr) = (3, 5)$ MIMO system for the selection of receiving antenna. Table 2 summarizes the observation from CDF plot. Figure 6 shows the BER plot for optimal selection and energy-aware selection.

The results discussed in Table 2 indicate that the maximum degradation is expected when the uncertainty is maximum. This has been indicated by the case when $Lr = 3$. The capacities for both selection criteria are indistinguishable when $Lr = 1$. Majority of the reported work had concentrated over the issue of antenna selection at transmitter side. A few contributions had reported different criteria for antenna selection over receiver side. Table 3 provides the state of the art of the proposed work.

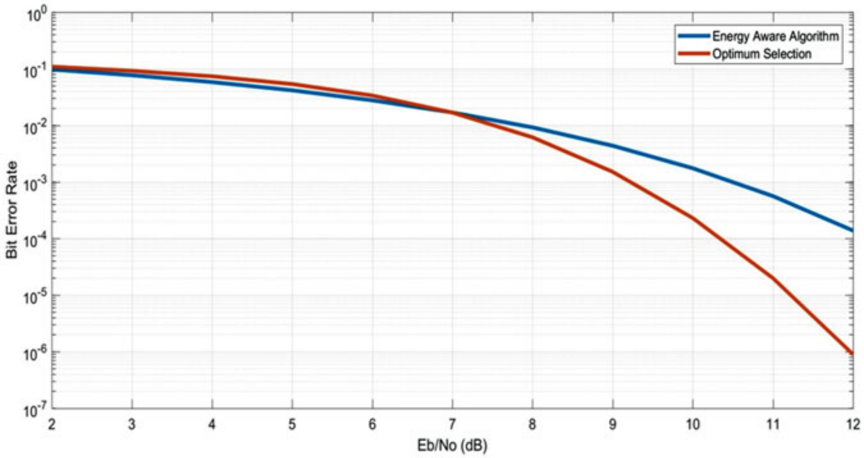


Fig. 6 BER analysis for optimal selection and energy-aware selection

Table 3 State of the art of the proposed work

Contribution	Operational side	Remark
Deotale et al. [3]	Transmitter	Selection process is complex due to orthogonal block coding
Berenguer et al. [4]	Transmitter	Selection process is complex due to discrete stochastic optimization
Zhang et al. [6]	Receiver	Selection process is complex cross-entropy optimization
Present work	Receiver	Selection process is simple as CDF of channel is used for decision

4 Conclusions

In this analysis, with the optimal selection method, as antenna positions are varied, a degradation from 25 bits to 21 bits are observed. Whereas, in energy-aware selection method a degradation of 0.5 bits per channel is observed. The selection of antenna is stuck up with the requirement of feeding mechanism and CSI calculation. The accurate selection is possible only when variation in the channel condition is slow. Hence, the present approach provides a simple and accurate way for the antenna selection in MIMO systems. In future, analysis for different coding technologies will be initiated for MIMO antenna selection.

References

1. Gesbert, D., Shafi, M., Shiu, D., Smith, P., Naguib, A.: From theory to practice: an overview of MIMO space-time coded wireless systems. IEEE J. Sel. Areas Comm. **21**, 281–302 (2003)
2. Sanayei, S., Nosratinia, A.: Antenna selection in MIMO Systems. IEEE Comm. Mag. 68–73 (2004)
3. Deotale, N., Kolekar, U., Hendre, V.: Ergodic capacity based antenna selection in MIMO for Rayleigh fading. In: International Conference on Information Proceedings (ICIP), Pune, India, pp. 317–321 (2015)
4. Berenguer, I., Wang, X., Krishnamurthy, V.: Adaptive MIMO antenna selection via discrete stochastic optimization. IEEE Trans. Signal Process. **53**(11), 4315–4329 (2005)
5. Yang, L.: MIMO systems with transmit antenna selection and power allocation over correlated channels. Wirel. Pers. Commun. **55**, 225–235 (2010)
6. Zhang, Y., Chunlin, J., Malik, W., Brien, D., Edwards, D.: Receive antenna selection for MIMO systems over correlated fading channels. IEEE Trans. Wirel. Commun. **8**(9), 4393–4399 (2009)

Public Auditing for Shared Data in Cloud Storage with an Effective User Dismissal

S. Samundiswary and Nilima Dongre

Abstract Cloud computing is an extensive technique which is changing the IT infrastructure swiftly. Data storage and sharing is the foremost and significant research area in cloud computing. Major security issues in cloud storage include missing folders, privacy settings, synchronization issues, etc. One of the topmost exigent research issues in data storage is data integrity. This research study identifies the security issue in data storage and achieves data integrity and privacy by providing public auditing using third-party auditor. To attain an effective key sharing, Shamir's secret sharing technique is exploited. The proposed system supports an efficient user dismissal by group admin in case if any user is found troublesome while sharing data in cloud storage. To provide a beneficial auditing for users, batch auditing is additionally presented to reduce the auditing time.

Keywords TPA · Cloud service provider · Proxy re-signature
Shamir's secret-sharing scheme

1 Introduction

Cloud computing provides many incredible benefits to end users. IaaS, SaaS and Paas utility services are the prime service recommendations to the client. Although cloud service providers offer many extraordinary benefits, still many organizations fail to implement it because of security concerns. The most used cloud computing service by users is cloud data storage, which minimizes the local data management and user overhead. Personal data, medical records and administration information demand a more secure place to store. In such cases, cloud storage fails to ensure data protection. Along with this; data breaches, cyberattacks, data loss, software and hardware vulnerabilities, malware injection and wrapping attack possibility are few

S. Samundiswary (✉) · N. Dongre
Department of Information Technology, Ramrao Adik Institute, Nerul, India
e-mail: s.samundiswary@yahoo.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_12

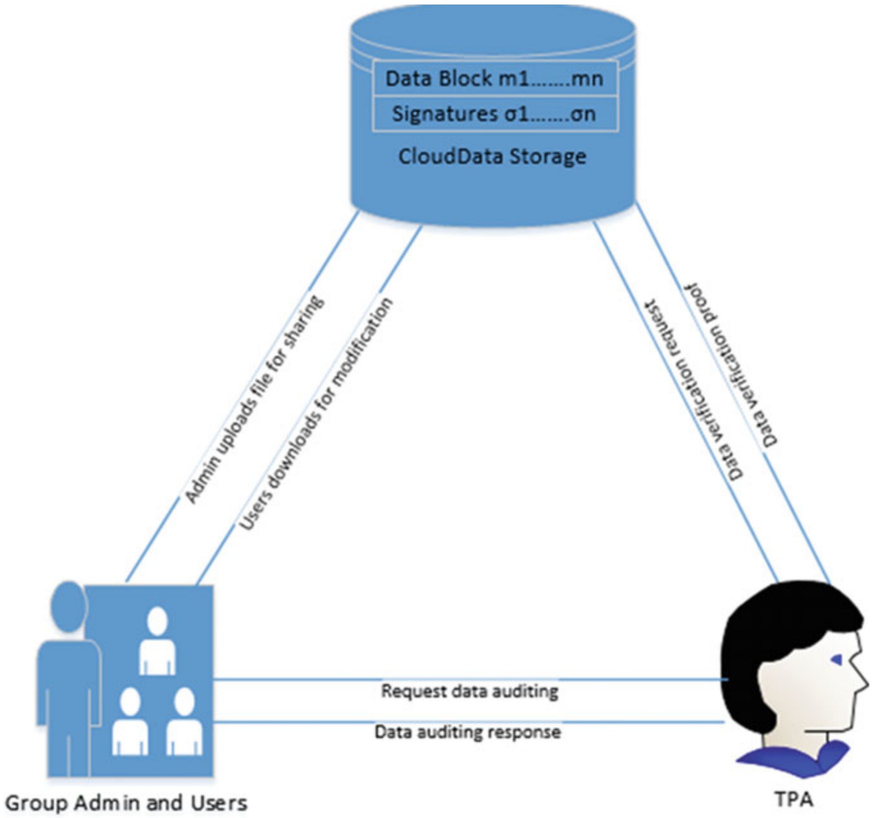


Fig. 1 Public auditing in cloud data storage

other concerns. There have been few cases, where some server faults with CSP have occurred in the past [1].

Data integrity in cloud storage is a pessimistic concern, since the possibility of external and internal data loss is supreme. Hence, data integrity must be validated in cloud storage ahead of doing any data manipulations like search, computation, etc. Some of the conventional data correctness technology makes use of downloading the whole data from cloud and checking the signature correctness like MD5 hash value, RSA signature, etc. Such techniques lead to expensive computation and resource cost (Fig. 1).

Many research works have been done to protect data integrity in semi-trusted cloud [2–7]. In these research works, a signature is attached with each data block. A common attribute considered in the referred work is public auditing, which is achieved using third-party auditor to reduce group admin responsibility. This research work not only includes public auditing but also considers how to maintain the identity of user from TPA during auditing process. Along with auditing and identity preservation, user dismissal is the most important attribute to be considered in data sharing. In this work, mischievous user is removed from the group because of security concerns. During this time, the removed user signature must be re-signed by group agent using proxy re-signature technique.

2 Related Work

Previous research works have provided public auditing technique to achieve secure cloud data storage as introduced by Cong Wang [2]. In this paper, author has proposed a protocol to achieve privacy-preserving public auditing using TPA as a verifier. To attain user's identity privacy, homomorphic linear authentication tag is used. This scheme also supports data dynamics like block-level insertion, modification and deletion. Another protocol was designed by Jilangto Li [3] as public auditing for low-performance end devices in cloud. In this, the author has concentrated auditing for devices like mobile phone and PDA's data in cloud using online and offline signatures. To design an online/offline signature, Chameleon hash function has been used. Since the author focused only on low-performance device, this protocol is not effective for large data auditing. Boyang Wang [4] introduced an alternative concept ring signature and provided a new dimension in public auditing of cloud storage with new feature of data sharing. To achieve block-less verifiability, the author recreated ring signature as homomorphism authenticable ring signature to attain user's identity preservation. Although the system attains block-less verifiability and non-malleability, it fails to attain user revocation. This system only focuses on static group of data sharing and as a result user cannot be added or removed from the group in between. To overcome ORUTA drawback, Boyang Wang designed a new public auditing system called KNOX [5], where the author applied homomorphic authenticable group signature to audit large group data in cloud. HAGS system consists of algorithms such as KeyGen, Join, Sign, Verify and Open. Finally, the author achieved storage correctness, identity privacy, large group support and traceability. Kai He [6] proposed a new protocol, which uses proxy re-signature concept mainly to attain user removal in group sharing. The author tried a user revocation without resigning revoked user signature. Users in the group are divided into pairs and each pair has a key called challenged key, which is preserved in CSP to use for re-signature in future user revocation. In this case, the author failed to focus on collision between revoked user and CSP. Jiawei Yuan proposed an integrity checking system with multi-user modification [7]. In this, the author applied new techniques like polynomial-based tag and proxy tag update techniques to avoid collusion between misbehaved user and cloud service provider. Instead of sending single-block tags for each user; multiple authentication tags are composed for all users and sent to the verifier for integrity proof verification. The security of the system is based on CDH, BDH and SDH technique. Recently, Tao Jiang [8] concentrated on avoiding collision attack and therefore ensured an efficient user revocation by utilizing technique vector commitment and verifier-local revocation group signature. This model does not work on data blocks; here vector commitment is applied on entire database of users to achieve zero-knowledge proof.

3 Proposed System

Although many public auditing systems have been designed previously, still an auditing system that resolves all integrity problems in cloud storage is yet to be designed. In this section, we provide a system, which satisfies public auditing with the help of TPA and privacy preservation by not disclosing the user information to TPA, considering the collusion resistance by utilizing single-use unidirectional proxy re-signature technique introduced by Giuseppe Ateniese [9] and Shamir's secret-sharing scheme [10] to share keys to users in the group.

3.1 System Outline

The proposed approach contains four main entities namely: Group admin, Group users, Third-party auditor and Cloud server. Data admin uploads the data to cloud server, which he wants to share with his group members. Along with the uploading of data, admin generates keys for accessing the data, which is shared amongst authorized users using Shamir's secret-sharing scheme. Admin dismisses the user when the user disobeys the rules of data sharing in group. Group user can access and modify data from CS. File uploaded by admin is split into blocks and stored on server. Such data blocks are downloaded and decrypted using separate private key of each user. User applies their own signature when the data is modified by them in CS. The TPA audits the data stored by users on server without seeing the information of data. The auditor only checks the hash value of data blocks instead of entire data. Cloud server is a major entity involved in this system, which helps in storing and sharing of user's data remotely. It stores keys and data separately to avoid attacks. Along with this, it contains proxy, which translates signature of dismissed user to the most existing user signature. This proxy internally uses unidirectional re-signature algorithm to achieve the above property. This technique avoids the collusion between dismissed user and cloud server.

3.2 Scheme Illustration

Key Creation: The group sharing public key is $P_{Ki} = g_{\xi i.}$ Whenever a user is dismissed from the group, a new public key is created by admin and distributed to all users.

User's private key is $S_{ki} = _{\xi i}$, which is distributed to all users through Shamir's sharing scheme. Group admin constructs a polynomial equation as follows:

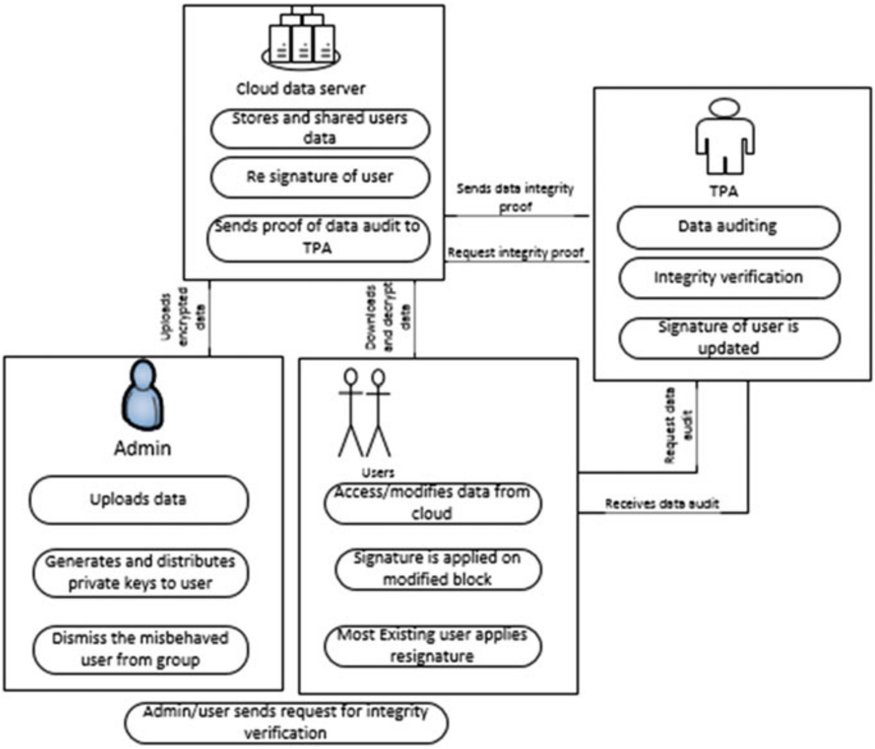


Fig. 2 Proposed public auditing system in cloud data storage

$$f(x) = s +_{a1x} + \cdots + {}_{ak-1}x^{k-1}, \quad (1)$$

where $s = 1/_{\xi i}$. The admin generates d random x^n and computes secret shares as $X_n = f(x_n)$ and sends (x_n, X_n) shares to all users in a group through secure channel.

Signature Creation: When user modifies blocks of data in a file, he/she must apply their own signature by using their own secret key on corresponding block as shown below:

$$\sigma_k = \left(h(id_k)\pi_{j=1}^{s}\left(g^{\alpha j}\right)^{mkj}\right)\xi_i, \quad (2)$$

where i_{dk} is block identifier of block $_{mk.}$ SHA 512 hashing technique is applied to generate signature of users (Fig. 2).

Rekey Creation: Cloud generates a random number R and sends to admin, admin then selects the most existing user from the user list and such user replaces the signature of revoked user. Most existing user u_c computes a value ξ_i/R and sends it. All users update the rekey in the group for further sharing of data. Proxy stores the keys of user in the group and verifies the equation.

$$e = (\sigma_1, g) = e(h(id_l)\pi_{j=1}^{s}\left(g^{\alpha j}\right)^{mkj}, p_{ka}). \tag{3}$$

Internally, the proxy uses proxy re-encryption technique to translate one user signature into another without the secret key. Signature of user is $\sigma_1^{(n)} = \sigma_1^{\xi n}$ and proxy sends the user's re-signature value(x_k, $\sigma_1^{(k)}$) to cloud server, where the re-signature of user (x_n, X_n) is converted to (x_k, $\sigma_1^{(k)}$).

Re-Signature Creation: Cloud server computes re-signature of revoked user after receiving from proxy. The re-signature of revoked user in block m_1 is updated. Finally, the revoked user signature is translated to the most existing user's signature.

$$\sigma_l' = \pi_k\left(\sigma_l^n\right)^{\lambda n 0 R}. \tag{4}$$

TPA Request: To audit the data, TPA sends challenge request message based on user request. For auditing, data blocks are sent to cloud.

$$CR = \left\{L, R, X^{\alpha j}\right\} 1 \le j \ge s\}, \tag{5}$$

where L = indices of selected blocks, R = random number and $X^{\alpha j}$ is value received from admin.

Cloud server Response-Cloud generates proof based on TPA request and sends it back to TPA.

$$Prf = \{wi\} 1 \le i \ge d, \pi\}. \tag{6}$$

TPA Proof verification: TPA compares the request equation and proof for checking integrity. It returns TRUE if equation holds true and if not it returns FALSE. It later sends it to group admin and users based on the request.

Batch Auditing:

TPA utilizes batch auditing method to audit many block requests from different users at the same time. Such batch auditing reduces time and increases the system efficiency. TPA sends batch TPA request to cloud server and cloud server replies back Batch Proof to TPA, which is again verified by TPA and then it is communicated to corresponding users and admin.

4 Experimental Analysis

In this segment, we have discussed the performance of our proposed design, which is implemented using JSP and Servlets. Group admin generates public key and private key using Java cryptography extensions (JCE). Private key is distributed to users using Shamir's secret sharing scheme. Each user uses the private key to update the blocks and applies hashing. Users who have public key are authorized to access the shared data. Whenever user revocation takes place, public key is newly distributed to avoid dismissed user to access the files. All experiments are tested under Windows 7 with 2.67 GHZ Intel R core(TM) i7 CPU and 4 GB RAM. The files shared to group users are text files. Performance of (i) Key generation time taken with respect to number of users (ii) Auditing time of TPA with respect to number of user requests (iii) Integrity verification, and finally comparison of batch auditing and single auditing with time is shown in Fig. 3, Table 1.

5 Conclusion

In this research work, group admin shares the data with authorized group members. To achieve data integrity in sharing data storage, public auditing system is designed newly. To avoid collusion between dismissed user and cloud server, single-use proxy

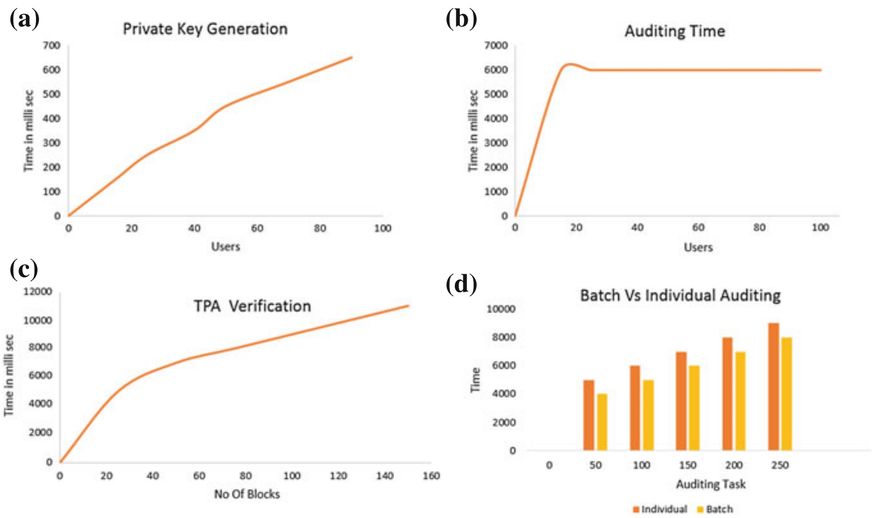


Fig. 3 **a** Key generation time **b** auditing time **c** verification time **d** comparison of individual and batch auditing

Table 1 Comparison of an existing auditing system and the proposed system in cloud data storage

Techniques	Existing system	Proposed system
Private key	No sharing technique	Shamir's secret-sharing scheme
Proxy re-signature	Bidirectional and re-signature applied by admin	Unidirectional method and applied by the most existing user
Public key	Single key for all users	Only to authorized users. Created newly after every user dismissal
Batch auditing	Available	Available

re-signature is utilized, and to reduce the admin workload TPA is introduced for auditing. The most existing user in the group re-signs when the user is dismissed from the group. Along with this above property, the identity of user is preserved from auditor, and SHA 512 hashing algorithm is applied by users to sign the data block. To improve an auditing time, our system also contains batch auditing mechanism. Public key of an authorized user is changed when any user is dismissed from the group. As a result, it restricts the dismissed user to again get involved in the sharing group.

References

1. Why are hackers increasingly targeting cloud. http://www.computing.co.uk/ctg/analysis/2430444/why-are-hackers-increasingly-targeting-cloud
2. Wang, C., Chow, S.S.M.: Privacy preserving public auditing for secure cloud storage. IEEE Trans. Comput. **62**(2) (2013)
3. Li, J., Zhang, L.: Privacy preserving public auditing protocol for low performance end devices in cloud. IEEE Trans. Inf. Forensic Secur. (2016)
4. Wang, B., Li, H.: Oruta: privacy preserving public auditing for shared data in the cloud. IEEE Trans. Cloud Comput. (2014)
5. Wang, B., Li, H.: Privacy preserving public auditing for shared cloud data supporting group dynamics. IEEE Trans. Cloud Comput. (2014)
6. He, K., Shi, C.H.J.: Identity preserving public auditing for shared cloud data. In: IEEE 23rd IWQOS (2015)
7. Yuan, J., Yu, S.: Efficient public integrity checking for cloud data sharing with multi user modification. In: IEEE INFOCOM (2014)
8. Jiang, T., Ma, J.: Public integrity auditing for shared dynamic cloud data with user revocation. IEEE Trans. Comput. (2015)
9. Ateniese, G.: Proxy resignature: new definitions, algorithms and applications. Springer J.
10. Yang, K., Jia, X.: An efficient and secure dynamic auditing protocol for data storage in cloud computing. IEEE Trans. Parallel Distrib. Syst. (2012)

Lightweight Effective Encryption Algorithm for Securing Data in Cloud Computing

Basel Saleh Al-Attab, H. S. Fadewar and Mahmoud E. Hodeish

Abstract Communication in the environment of cloud computing is implemented through the Internet and its backbone. Some issues of network security in the cloud environment are caused by its essential characteristics such as resource pooling, virtualized nature, elasticity, and other measured services. Though many algorithms have been used to secure the data communication in the cloud environment, some problems in the use of such algorithms still exist. Some of such problems are the mathematical complexity, key and security weakness, time complexity, and slow performance. In this paper, an algorithm called hyper data encryption (HDE) is proposed to combine the symmetric ciphers, secret sharing, and Diffie–Hellman key exchange concepts in order to enhance the security and solve the mentioned problems. Performance analysis is conducted in terms of key-space analysis, key sensitivity analysis, correlation analysis, information entropy analysis, time complexity, and execution time. The results show that the proposed algorithm is better in a cloud environment, which can provide strong security and high performance.

Keywords Cloud computing · Cryptography · Secret sharing · Key exchange

1 Introduction

Cloud computing is one of the most recent core technical topics in the network security field. It refers to a model for enabling resource pooling, convenient, and ubiquitous, which can be easily delivered by various types of service that provide interaction based on demand access [1]. Cloud computing is suitable for accessing

B. S. Al-Attab (✉) · H. S. Fadewar · M. E. Hodeish
School of Computational Sciences, S.R.T.M University, Nanded, India
e-mail: atbbasel@gmail.com

H. S. Fadewar
e-mail: fadewar_hsf@yahoo.com

M. E. Hodeish
e-mail: mah_hodeish@yahoo.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_13

data anywhere at any time and it has an ability to handle the increasing online users on networking sites, video conferencing, and online surfing. In other words, cloud computing has eliminated the necessity for clients to be present at the software, hardware, and storage space located. Thus, cloud computing provides a flexible environment for data exchange [2]. Since the communication in cloud computing environments is via the Internet and its backbone, security requirements have become an essential task. There are many security issues, which are the mathematical complexity, key and security weakness, time complexity, and slow performance. Thus, several security issues of the technologies and technology systems mentioned above are appropriate to be used for the cloud computing [3]. Network security is considered as one of the major security concerns of cloud computing, which is concerned with both internal and external attacks [4, 5]. It provides a secure communication between users and cloud servers. Having a secure network is necessary to transmit a secure data [6]. To ensure the security of data exchange over networking, many technologies and techniques have been proposed, where cryptography is the most efficient. Cryptography includes the conversion of plaintext into ciphertext. Frequently, it is a technique used to transfer contents safely by ensuring that only the intended receiver can discover them [7].

The paper is organized as follows: the related work is discussed in Sect. 2 and in Sect. 3, the preliminary notations are defined. The proposed method is explained in Sect. 4 and in Sect. 5 the experimental results are shown. The performance analysis is discussed in Sect. 6. Finally, the paper is concluded in Sect. 7.

2 Related Work

Many cryptographic algorithms have been proposed to provide security for cloud computing. This section presents a review of some popular algorithms and some of the recent studies that investigated the cryptography for cloud computing. The popular cryptographic algorithms are categorized as advanced encryption algorithm (AES), data encryption standard (DES), and international data encryption algorithm (IDEA). DES is a commonly used symmetric key algorithm invented in 1974 by IBM. Many proposed algorithms have recently proven that the DES algorithm is unsecured [8]. DES used small key size which results in weak security, easily broken. In addition to this particular drawback, DES is also observed to work slowly on software [9]. AES is declared as O.S. FIPS NIST in 2001, an unbreakable algorithm with high efficiency. However, AES has some shortcomings such as it needs more processing and requires more rounds of communication as comparing to DES. Regarding the IDEA, James L. Massey of ETH Zurich and Xuegialai designed the block cipher algorithm IDEA, which has been introduced in 1991. It is worth reminding that the algorithm came through some adjustments to be known later as IDEA. In fact, IDEA consists of a high number of weak keys, which is considered as its main drawback. In addition, a new attack on round six of IDEA is detected [10]. In a recent investigation on cryptography for cloud computing, cryptography for cloud computing and cloud security on the base of public key values has been tackled by many contemporary scholars such as Lyer and Sanyal [11]. They discussed an AES encryption algorithm with multi-key (128, 192, 256) bit cipher key used to cipher and decipher data. Their study confirmed that the security of AES is better than RSA security for cloud computing. Meanwhile, AES has an ability to be used in public and private clouds as well as in the virtual machine. In this regard, the main idea for a multivariate key cryptography and homomorphic encryption has been proposed by Ustimenko and Wroblewska [12]. They found that algebra has an important role in cryptography in order to secure the cloud computing security. Several researches on cloud computing security have investigated the secure paths of cryptography such as data integrity and privacy [13]. Such studies only have examined the concealment of information from clients and users. Therefore, a secure way is based on advanced AES encryption for hiding information sessions between clients and servers to construct cloud computing platforms [13]. In this model, the asynchronous key system and AES-based file encryption systems for exchanging data or information are included. Atyero and Feyisetan [14] have explored the secure delivery of data sessions from the cloud and noted serious issues on such delivery. In order to address serious security concerns in terms of access to data cloud, they proposed the use of homomorphic encryption. Regarding symmetric and asymmetric algorithms, Bhardwaj et al. [15] have emphasized on symmetric algorithms for security consideration on which one should be used for cloud computing environment that requires data and links encryption. They examined the symmetric algorithms of various encryption techniques and concluded that the AES is a better candidate for key encryption. Kaushik et al. [16] have proposed a hybrid symmetric cryptography process for protecting data stored at cloud server from any malicious activity, on which it combines substitution cipher and transposition cipher techniques for the encryption and decryption of data. Anyway, some of the cryptographic algorithms have weak security and others require high processing. Therefore, there is a dire need to design an algorithm to be proper for cloud computing in terms of security and processing. In this study, hybrid data encryption (HDE) algorithm is proposed to serve the security of cloud computing with high performance and low processing. It is a hybridization between symmetric key, secret sharing, and key exchange.

3 Preliminary Notations

3.1 Secret Sharing

In secret sharing schemes, the secret is spirited into a number of shadows and distributed among n participants. Where any ($k <= n$) stacks their shadows together the secret can be recovered. A visual secret sharing (VSS) proposed by Naon and Shamir [17] is a special method of secret sharing. It is used to encrypt the secret message in visual form by splitting it into n shares, then it will be transmitted securely via the

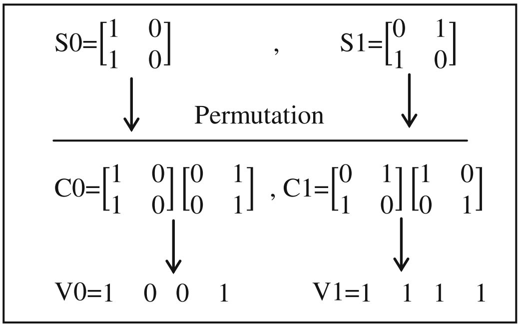


Fig. 1 Basis matrices of VSS

Internet or any other communication channels. The secret message can be decrypted when a sufficient number of shares are stacked together using OR of XOR operation. The proposed method is to use the concept of VSS in encryption and decryption purpose based on XNOR operation. To illustrate the idea behind the concept of VSS [18], 2-out-of-2 scheme (2 subpixels) is explained in Fig. 1.

The mentioned scheme is applied on binary-valued images with values belonging to [0, 1]. A binary image pixel is divided into two subpixels, out of which white (1) or black (0) is randomly chosen depending on the current pixel as shown in Fig. 1 in this scheme. Choose one of the two rows of white, if the image pixel is white, or choose between one of the two rows for black, if it is black. A random column permutation of the white pixel and the black pixel is done from S_0 to S_1 to generate a given matrix as C_0 and C_1 that represents white and black pixel, respectively, which produces two vectors V_0 and V_1 corresponding to the occurrence of the pixel, i.e., either white or black in a secret image as shown in Fig. 1. If the pixel is white then V_0 will be outputted, which is of either *1 0* or *0 1* value with gray-level ½ and if the pixel is black then V_1 will be outputted with either *1 1* or *1 1* value.

3.2 Diffie–Hellman Key Exchange

The Diffie–Hellman (D–H) key agreement algorithm is a method that allows two parties (devices and users) to communicate over a network by establishing a shared secret key without exchanging any secret data. The security of this algorithm is based on solving discrete logarithmic problem. Suppose that Alice and Bob are to agree on a session key over insecure networks. The parameters g and p are public.

3.3 Transpose of Matrices

The way of turning all the rows of a given matrix into columns and vice versa forms a matrix. The transpose of a matrix A is expressed as A^T. Transpose matrices are proper in cryptography because it helps in restoring the original data after being encrypted. This can be proved through the following transpose matrices property:

If A represents a matrix, then

$$(\mathrm{A}^{\mathrm{T}})^{\mathrm{T}} = \mathrm{A}. \tag{1}$$

Equation (1) shows that the transpose of a transpose matrix is the original matrix. To illustrate that, the following example is given:

$$\mathrm{A} = \begin{bmatrix} A & B & C & D \\ E & F & G & H \\ I & J & K & L \\ M & N & O & P \end{bmatrix}, \mathrm{A}^{\mathrm{T}} = \begin{bmatrix} A & E & I & M \\ B & F & J & N \\ C & G & K & O \\ D & H & L & P \end{bmatrix}, (\mathrm{A}^{\mathrm{T}})^{\mathrm{T}} = \begin{bmatrix} A & B & C & D \\ E & F & G & H \\ I & J & K & L \\ M & N & O & P \end{bmatrix}.$$

4 The Proposed Method

In order to enhance the security of cloud computing with high performance and low processing, the new algorithm HDE is proposed with hybridization between symmetric key, secret sharing, and key exchange, where the key exchange technique is used to generate a key, and the secret sharing technique is used for encryption and decryption procedures with aspect of the symmetric key technique. HDE consists of four procedures namely key generation, secret sharing, transpose, and swap, which are working together for encryption and decryption purposes. The detailed steps of the procedures are described as follows.

4.1 Key Generation Procedure

The key generation algorithm is based on D–H algorithm in order to enable the user and cloud exchange key that can then be used for subsequent symmetric encryption of messages securely. The proposed algorithm has modified the D–H algorithm to generate and exchange a key matrix with the size of 4×4 rather than exchanging single key in the classical D–H algorithm. It is to say that this is one of the contributions of this paper. In addition, the private key of user and cloud is generated randomly based on the concept of the master random grid as a matrix with the size of 4×4.

4.2 Secret Sharing Procedure

The concept of visual secret sharing applied on image data is used in the proposed method, whereas in this method, it is applied on text data with (2, 2) access structure without expansion of data. The basic matrices of this procedure are produced on the base of XNOR operation that will be used in the encryption procedure.

$$C_0 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, C_1 = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}.$$

Here, each word (W) of each block will be converted into binary form (8-bits) with values belonging to [0, 1]. Then through the above matrices, each bit will be split into two sub-bits randomly according to visual secret sharing concept. After all the bits of the word being scanned, two shares of the secret will be generated in binary-valued forms.

- Procedure

1. Take the first block of the message.
2. For each block, take a first word (W) which is converted into 8-bit binary.
3. For each bit, determine its value:

a. If *0*: Randomly select any row from C_0. Randomly assign one element of the selected row into the first share and the other into the second share.
b. If *1*: Randomly select any row from C_1. Randomly assign one element of the selected row into the first share and the other into the second share.
4. Repeat Step 3 till all words of block shared.
5. Repeat Steps 1 to 4 till all blocks of messages shared.
6. Convert the share 1 and share 2 into their original form.

Cryptographically, confusion and diffusion are two properties of the operation of a secure cipher. Confusion can be used drastically to change the data from the input to the output. Whereas diffusion is used to make a change in many characters of the output by changing a single character of the input. Two procedures, called transpose and swap, are used in the proposed algorithm in order to make such confusion and diffusion for the data and key.

4.3 Transpose Procedure

Apply the transpose of matrix operation on the input matrix data by turning all of its rows into columns and vice versa using Eq. (2).

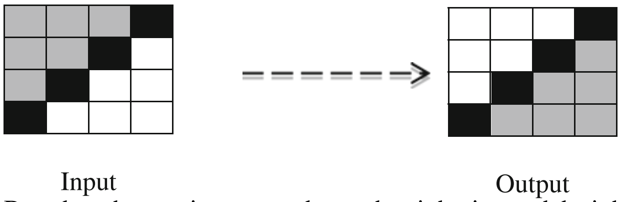
$$\text{Output} = (\text{input})^{T}, \tag{2}$$

where input indicates the input matrix data and output indicates the output matrix data.

4.4 Swap Procedure

Take the input data matrix by swapping all the upper triangular entries above the main diagonal with the lower triangular entries below the main diagonal as follows:

$$\forall aij \in A : swap(aij, aji); i \neq j, i, j = 1, \ldots n. \tag{3}$$



Based on the previous procedures, the ciphering and deciphering processes works as follows:

4.5 Encryption Procedure

1. Divide D the secret message M into blocks B, each block with the size of 16 bytes (4×4 matrix).

$$\mathrm{D(M)} \rightarrow \mathrm{B_n} : \forall \mathrm{B_i} = \mathrm{B_i}\,[4, 4]\&\mathrm{B_i} = 16\,\mathrm{byte}, \quad \mathrm{i} = 1, \ldots, \mathrm{n}.$$

Pass each block to the secret sharing procedure SH in order to produce two shares of secret messages (secret 1 and secret 2).

$$\forall B_i, B_i \rightarrow SH(B_i) = (secret\,1\,\&\,secret\,2),\ i = 1, \ldots, n.$$

Take the generate key K of key generation procedure and perform XNOR operation between K and the two shares of secret messages as follows:

$$\begin{aligned} S_1 &= \mathrm{K} \otimes \mathrm{secret}\,1 \\ S_2 &= \mathrm{K} \otimes \mathrm{secret}\,2, \end{aligned} \tag{3}$$

where $\otimes$ denotes the XNOR operation.

Apply the transpose procedure on the S_1 and S_2. Here, the S_1 and S_2 are represented as two matrices with the size of (4 × 4). The transpose operation is used to make the rows of the original matrix as columns of the new matrix and vice versa.

$$ST_1 = Transpose\,(S1)$$
$$ST_2 = Transpose\,(S2). \tag{4}$$

Take the key into swap procedure

$$K_1 = swap\,(k). \tag{5}$$

Take ST_1 *and* ST_2, then perform XNOR operation between them and K_1 in order to produce the final encrypted shares of the secret.

$$\mathrm{Final} - \mathrm{S}_1 = ST_1 \otimes K_1$$
$$\mathrm{Final} - S_2 = ST_2 \otimes K_1. \tag{6}$$

Repeat the previous steps from 2 to 6 till all blocks of the secret message are being encrypted.
Transmit the final shares of encrypted messages into user/cloud through one or two secure channels.

Figure 2 shows the overall structure of the HDE encryption procedure.

4.6 Decryption Procedure

Apply the reverse process of the encryption procedure, but at the secret sharing process just perform the XNOR operation between the two shares of secret (*secret 1 and secret 2*) in order to recover the original message.

$$\mathrm{Plaintext} = secret\,1 \otimes secret\,2. \tag{7}$$

5 Experimental Results

Experimental results are conducted in this section to present and verify the feasibility of the proposed method. The experiment is conducted on text data type with 128 bits key size and the block size is 128 bits. In key generation procedure, the key with the size of 128 bits, which is formulated as a 4 × 4 matrix is generated based on the modified D–H key exchange. Key exchange is based on the case of the prime number q = 353 and a primitive root of 353, in this experiment α = 3. User and cloud

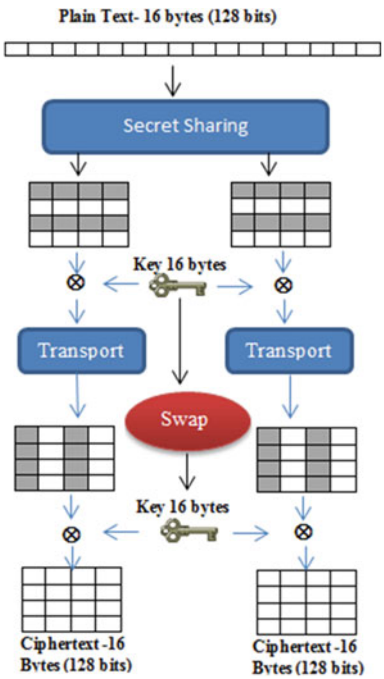


Fig. 2 HDE encryption procedure

generate a private key matrix with size of 4×4 based on the RG concept as shown below:

$$\text{User A} = \begin{bmatrix} 291 & 206 & 289 & 150 \\ 347 & 38 & 92 & 110 \\ 258 & 320 & 210 & 57 \\ 121 & 311 & 7 & 63 \end{bmatrix} \quad \text{Cloud B} = \begin{bmatrix} 149 & 246 & 24 & 144 \\ 33 & 247 & 113 & 190 \\ 211 & 226 & 187 & 254 \\ 166 & 11 & 231 & 342 \end{bmatrix}.$$

Each of them has to compute its public key. User computes YA = 3 (user A) mod 353, and cloud computes YB = 3 (cloud B) mod 353.

$$\text{YA} = \begin{bmatrix} 107 & 259 & 208 & 78 \\ 215 & 43 & 128 & 293 \\ 186 & 58 & 152 & 105 \\ 10 & 45 & 69 & 247 \end{bmatrix} \quad \text{YB} = \begin{bmatrix} 26 & 334 & 256 & 295 \\ 67 & 296 & 145 & 271 \\ 103 & 198 & 59 & 303 \\ 18 & 244 & 106 & 335 \end{bmatrix}.$$

After they exchange public keys, each can compute the common secret key. User computes k = (cloud B) user A mod 353, and cloud computes k = (user A) cloud B mod 353.

Table 1 An example of data after being encrypted in HDE

Text	ABCDEFABCDEFABCD
Key	143 50 86 132 132 72 78 150 91 148 110 132 19 1 98 27
HDE 1	´Eå ª!UJ,7]~Õ¡È
HDE 2	I × 8÷ ¼(÷:ñjÀhœ

$$K = \begin{bmatrix} 297 & 352 & 182 & 337 \\ 59 & 38 & 177 & 70 \\ 153 & 231 & 49 & 76 \\ 304 & 101 & 286 & 127 \end{bmatrix}.$$

The key generated based on exchange process is used to encrypt the data in the encryption procedure. Three subprocedures are included in the encryption procedure as secret sharing, transpose, and swap. Suppose that the data of such block data with key is shown in Table 1 along with the HDE 1 and HDE 2.

The block data is taken into secret sharing procedure in order to produce two ciphers of data. Each word (w) should be converted into 8-bits binary. For example, the w = "A", and the 8-bit binary values after being converted are as follows: A = 01000001. Now randomly two shares of binary values belonging to [0, 1] will be generated as follows:

Share 1 Share 2	00110011 10001101
Share1⊗ Share 2	01000001 ———→sel1 "A"

After finishing the remaining steps of the encryption algorithm, two ciphers will be generated as shown in Table 1. By reversing the steps of the encryption procedures the original data is recovered. Briefly, in HDE algorithm, size key of 128 bits is used. HDE works on 4 × 4 matrices. It consists of key exchange, secret sharing, matrices transpose, and swap. Due to the use of one round, it works fast on both software and hardware.

6 Performance Analysis

Various analyses are conducted to measure the performance of the proposed method with comparison to different cryptographic algorithms such as DES, AES, and IDEA.

Table 2 A form of data after being encrypted and decrypted in scheme and the compared algorithms

Cryptography technique	Original secret	Encrypted data	Decrypted data
DES	ABCDEFABCDEFABCD	++sC|r Ã Pˆ « ?{A	ABCDEFABCDEFABCD
AES	ABCDEFABCDEFABCD	÷üjðK?hTMØ	ABCDEFABCDEFABCD
IDEA	ABCDEFABCDEFABCD	Ôî+±0 m{WyQ%	ABCDEFABCDEFABCD
HDE 1	ABCDEFABCDEFABCD	´Eå ª!UJ,7]~Õ¡È	ABCDEFABCDEFABCD
HDE 2	ABCDEFABCDEFABCD	Ix8÷¼(÷:ñjÀhœ	ABCDEFABCDEFABCD

Table 3 Key size and space of HDE and the compared algorithms

Cryptography technique	Key size	Key space
DES	64	264
AES	128	2128
IDEA	128	2128
HDE	128	2128

6.1 Visual Assessment

It needs to be performed on the encrypted output. If an attacker has an ability to deduce some meaningful information by visual assessing the encrypted result, then the scheme is said to be abortive in the first step itself. So, for a scheme to be successful, the attacker should not be able to deduce any meaningful information out of the encrypted result. As encrypted output shown in Table 2, a hacker has an ability to deduce some meaningful information by visual assessing the encrypted result. It is to say that the proposed algorithm is successful due to that and there is no relationship between the original secret and the encryption result.

6.2 Key-Space Analysis (Brute-Force Attack)

Encryption schemes should be highly sensitive toward very small change in the key used during encryption. Using a large key space ensures resistance of the technique toward brute-force attacks, which will need a maximum complexity of $2^{key-size}$ to find the correct key. Table 3 shows that the key size and key space of the proposed algorithm, AES, and IDEA are the same, but the generated key in the proposed algorithm is key exchange based on the modified D-H algorithm. In addition, the master RG is used to make randomness. So, there is no chance for brute-force attacks to break the proposed algorithm.

6.3 Key Sensitivity Analysis

Key sensitivity analysis is performed with the aim to check the sensitivity of encryption scheme toward change in initial conditions. It means that a slight change in the encryption key should produce an entirely different ciphertext. Due to usage of the randomness with the D–H algorithm to generate the key, it is proved that the proposed algorithm is more sensitive key than the compared algorithms.

6.4 Statistical Analysis

This analysis is done to analyze the confusion properties of an encrypted data. Evaluation of correlation coefficients shows the relation between the encrypted data and original data. Correlation analysis can specify the technique that has better confusion properties and can resist statistical attacks.

6.4.1 Correlation Analysis

This analysis calculates the correlation among the encrypted data and original data. A good encryption technique should result in an encrypted data with zero correlation ideally. In order to calculate the correlation between plaintext, the following formulae should be used:

$$r\alpha\beta = \frac{\mathrm{cov}(\alpha, \beta)}{\sqrt{\mathrm{D}(\alpha)}\sqrt{\mathrm{D}(\beta)}}$$

$$cov(\alpha, \beta) = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} (\alpha \mathrm{i} - \mathrm{E}((\alpha))(\beta \mathrm{i} - \mathrm{E}(\beta))$$

$$D(\alpha) = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\alpha\mathrm{i} - \mathrm{E}(\alpha))^2$$

$$D(\beta) = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\beta\mathrm{i} - \mathrm{E}\ (\beta))^2, \tag{8}$$

where α *and* β denote two values for which correlation needs to be calculated, N is the total number of elements obtained from the data, whereas $E(\alpha)=mean$ of α, and $E(\beta)=mean$ of β. As shown in Fig. 3, the IDEA has a small value of correlation coefficient, while the HDE is better as compared to DES and AES.

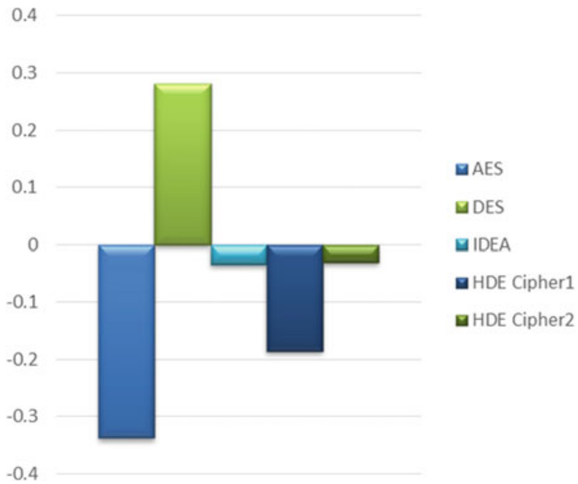


Fig. 3 Graphical representation of the correlation analysis

6.5 *Information Entropy Analysis*

The amount of randomness can be measured by information entropy. Mathematically, the entropy H(S) of message m can be expressed as follows:

$$H(S) = \sum_{i=0}^{n-1} \mathrm{P(S_i)log_2} \frac{1}{\mathrm{P(S_i)}}, \tag{9}$$

where $P(S_i)$ indicates the probability symbol S_i, log is of base 2. If there are 256 possible outcomes of the message S with equal probability, the H(S) = 8, which is a standard value for this case. The value of entropy close to value 8 indicates that encrypted output is highly random in nature. The randomness in encrypted data is measured by entropy. In comparison to the method of Entropy, the HDE provides satisfying results that come close to the ideal value 8 as shown in Fig. 4.

6.6 *Time Complexity*

For the key size of 128 bits, an attacker needs to find 2^{128} possible keys. Thus, the time complexity of 2^{128} complexity of finding the correct key is of $\boldsymbol{O}$ (1) on average. Indeed, the time complexity of the HDE is equal to the AES, but it has high performance due to the reason that there are no more iterations as in AES and the rest of the compared algorithms.

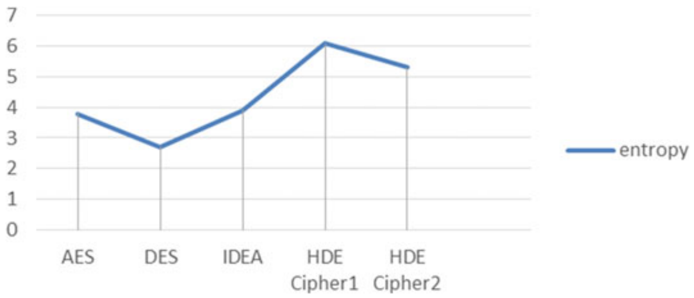


Fig. 4 Graphical representation of the entropy analysis

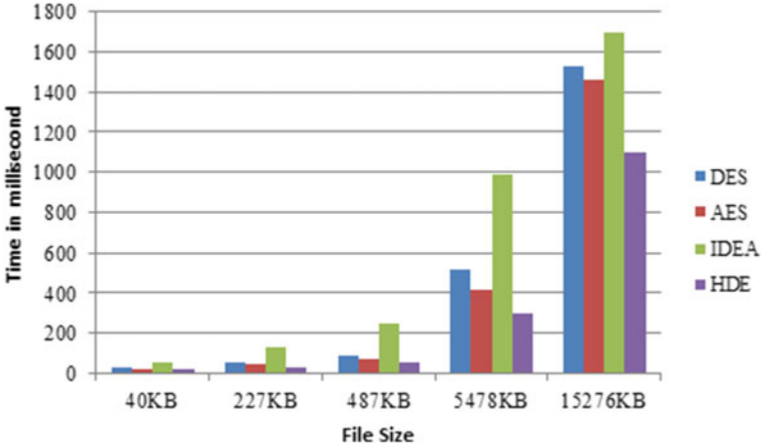


Fig. 5 Execution time of encryption for different file sizes

6.7 Execution Time

Execution time is the time, which will be taken by the algorithm for the encryption and decryption of file. Figure 5 shows the execution time of encryption in milliseconds for the compared algorithms with the respect of different sizes of file. Figure 6 shows the execution time of decryption in milliseconds for the compared algorithms with the respect to different sizes of file. It is clearly shown that the HDE takes less time than the DES, AES, and IDEA for encryption and decryption, i.e., HDE is better than the other algorithms.

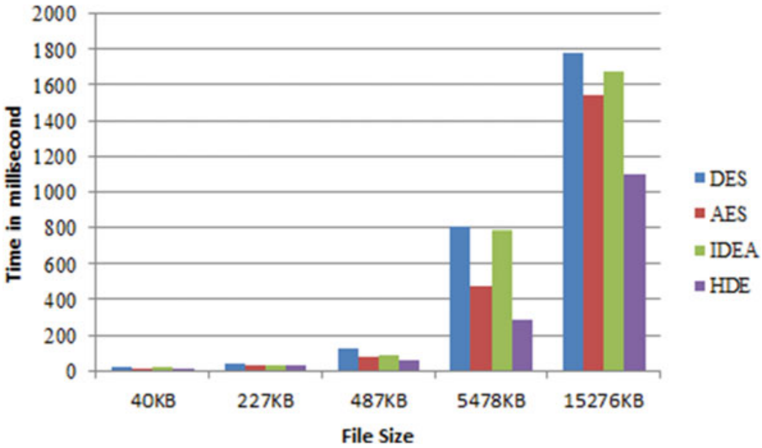


Fig. 6 Execution time of decryption for different file sizes

7 Discussion

In this section, we present a brief discussion of our proposed algorithm based on security features of data encryption in cloud computing.

- Key Space: HDE uses a large key space, which helps in resisting the brute-force attacks because of exchanging a key in matrix form rather than exchanging a single key as used in traditional D–H. Since the key exchange method is used, the security level will be increased.
- Time Complexity: There is no large time complexity due to the removal of the requirements of more rounds and iterations of the symmetric key methods.
- Security: HDE is a secure algorithm, because of using a complex structure such as secret sharing, transpose, and swap procedures. As a result of merging between secret sharing and symmetric key concepts, the proposed method has an ability to offer a high security with no minimum correlation between the input and output.
- Reliability: Secret sharing made the proposed algorithm more reliable and secure due to producing two shares, where each share can be known as a component of data, which contains incomplete information about the original data. Hence, it seems like meaningless information.
- Integrity: Due to the use of the transpose and swap procedures, a small change in input data will bring a drastic change in the enciphered output.
- Storage: The proposed algorithm is useful for distributed storage system in cloud computing environment because of using the secret sharing concept in order to provide a reliable access to data over individually unreliable nodes.

8 Conclusion

Data and network security is an important problem in cloud computing environment that needs to be firmly encrypted. Therefore, cryptography is one of the efficient techniques to ensure the security of data. In this study, hybrid data encryption (HDE) algorithm is proposed, which includes using the concepts of symmetric key algorithm, secret sharing, key exchange, and mathematical algebra. Therefore, it enjoys all the features of its components to be proper for the security of cloud computing. Performance analysis with comparison to some popular cryptographic algorithms like DES, AES, and IDEA is offered. Moreover, the analysis has proved that the proposed algorithm is robust, secure, and sensitive with respect to the key length and randomness. In addition, it is more efficient for cloud computing environment with respect to its rapid processing data and processing time. The future work is to enhance and develop an algorithm for cloud security with respect to the lightweight cryptography features.

References

1. Deshpande, P., Sharma, S.C., Sateesh Kumar, P.: Security threats in cloud computing. In: International Conference on Computing, Communication & Automation (ICCCA), pp. 632–636 (2015)
2. Pansotra, A., Preet Singh, S.: Cloud Security Algorithms. Int. J. Secur. Appl. **9** (10) (2015)
3. Saleh, A.-A.B., Fadewar, H.S.: Security issues and challenges in cloud computing. IJESE **2**(7), 2319–6378 (2014)
4. Deshpande, P., Sharma, S.C., Peddoju, S.K., et al.: Security and service assurance issues in Cloud environment. Int. J. Syst. Assur. Eng. Manag. (2016). https://doi.org/10.1007/s13198-016-0525-0
5. Deshpande, P., Sharma, S.C., Peddoju, S.K.: Data Storage Security in Cloud Paradigm. In: Proceedings of Fifth International Conference on Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing, vol. 436. Springer, Singapore (2016)
6. Daya, B.: Network Security: History, Importance, and Future University of Florida Department of Electrical and Computer Engineering
7. Jaber, A.N., Fadli, M., Zolkipli, B.: Use of cryptography in cloud computing. In: IEEE International Conference on Control System, Computing and Engineering, 29 Nov–1 Dec 2013
8. Thakur, J., Kumar, N.: DES, AES, blowfish: symmetric key cryptography algorithm simulation based performance analysis. Int. J. Emerging Technol. Adv. Eng. (IJETAE) **1**(2) (2011), ISSN: 2250-2459
9. Pansotra, A., Preet Singh, S.: Cloud security algorithms. Int. J. Secur. Appl. **9**(10), pp. 353–360 (2015)
10. Patil, S., Bhusari, V.: An enhancement in international data encryption algorithm for increasing security. Int. J. Appl. Innov. Eng. Manag. **3**(8), Aug 2014
11. Sanyal, S., Iyer, P.P.: Cloud Computing-An Approach with Modern Cryptography (2013). arXiv:1303.1048
12. Ustimenko, V., Wroblewska, A.: On some algebraic aspects of data security in cloud computing. In: Proceedings of Applications of Computer Algebra ACA 2013, Málaga, p. 155 (2013)
13. Wazed Nafi, K., Shekha Kar, T., Anisul Hoque, S., Hashem, M.: A Newer User Authentication, File encryption and Distributed Server Based Cloud Computing Security Architecture (2013)

14. Atayero, A.A., Feyisetan, O.: Security issues in cloud computing: the potentials of homomorphicencryption. J. Emerg. Trends Comput. Inf. Sci. **2**(10), 546–552 (2011)
15. Bhardwaj, A., et al.: Security algorithms for cloud computing. Procedia Comput. Sci. **85**, 535–542 (2016)
16. Kaushik, S., Gandhi, C.: Cloud data security with hybrid symmetric encryption. In: 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT). IEEE (2016)
17. Naor, M., Shamir, A.: Visual cryptography, advances in cryptography-eurocrypt '94, vis. Lect. Notes Comput. Sci. **950**, 1–12 (1994)
18. Hodeish, M.E., Humbe, V.T.: A (2, 2) secret sharing scheme for visual cryptography without Pixel Expansion. In: 2015 IEEE International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO) (2015)

Predictive and Prescriptive Analytics in Big Data Era

Prachi Deshpande

Abstract The notion of data analytics and its real-time application is important in the big data era owing to the voluminous data generation. Predictive and prescriptive analytics provides the future trends from the available data effectively. This will help to decide the usability of the data and thereby its retention for future applications. The paper reports the predictive and prescriptive analytics notion in big data regime, various platforms for its analysis, and the future research directions.

Keywords Analytics · Big data · Data science

1 Introduction and Driving Force

The present era is of information explosion with Internet of Things (IoT) and very soon the Internet of Everything (IoE) will ensure the connectivity of millions of gadgets and devices to the Internet [1]. According to [2], today 2.5 Exabyte (10^{18}) of data is created every day and 90% of the data is created in the last 2 years. Such an enormous data, generated from various user-driven applications, paved the way for many new avenues across different fields starting from household to business applications. Academicians and industry referred such a huge data with a notion of big data.

The term *big data* was first announced by *Cox and Ellsworth* in 1997 [3] in the context of data handling and processing capacity of available computers. Soon the term *big data* became the catchphrase for the academicians and industry. Big data can be characterized by five Vs as shown in Fig. 1.

Increased social media applications have increased the volume as well as the variety of data generated. It is expected to have 5 million mobile phone users in the world. The biggest challenge in front of big data is its processing and veracity. Only real-time processing of such a voluminous data will help to capitalize the usability

P. Deshpande (✉)
Nanded, India
e-mail: deprachi3@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_14

Fig. 1 Big data characteristics

of the data in the decision-making process. Value of the generated data is another important concern of big data. To increase the maximum gain out of the generated data, its analytics in different forms are required.

1.1 Classification of Data in Big Data Era

The data generated from different sources may differ in terms of its scale, distribution, domain, representation, and its density. Figure 2 depicts the classification of data in big data era.

Data science technology had defined, categorized, and processed such a data in efficient manner. Data science is an interdisciplinary field, which extracts meaningful information from huge data by using scientific techniques, methods, and processes. With the advent of data science technology, it becomes easy to predict the behavior of users. This aspect is known as data analytics. The increased use of social media has given rise to the big data in terms of online social network data.

Mobile and IoT data is another source of big data. With the emergence of 5G technology, the analysis of such data becomes an important concern for data scientist/analytics. With the advent of the concept of smart cities, the online social media data will very soon consist of geographical data in real-time mode. This will shift the research paradigm toward building of 3D geographic database of the locations. This advancement will shift the data science to a new direction of spatiotemporal data, the data with many features for selection and analysis. The spatial big data may consist of vector, raster, and networked data. The advancement in the spatiotemporal

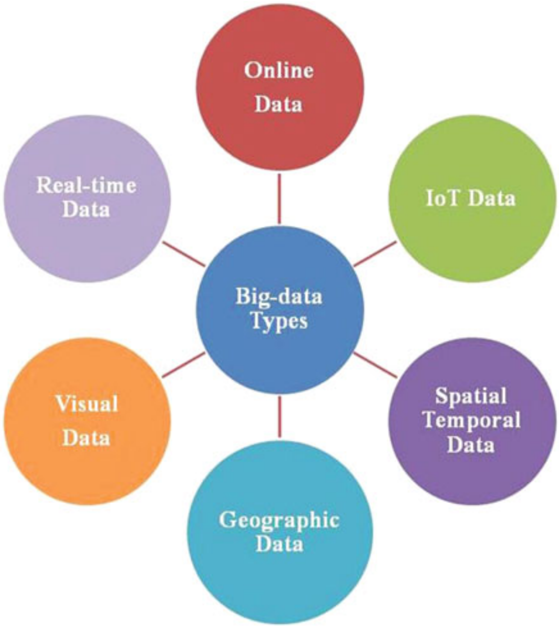


Fig. 2 Data types in big data Era

Table 1 Normal data and big data

Parameter	Normal data	Big data
Volume	In Gigabytes (GB)	In Tera (TB) or Peta (PB) bytes
Data management	By RDBMS	Hadoop, MapReduce, NoSQL, and high-performance computing
Growth interval	Measured in hours or daily basis	No such fixed scale for measurement
Data format	Structured	Structured, semi-structured, or unstructured
Data access	GUI-based interactive	Batch processing or real-time system

data will give rise to the notion of real-time streaming data due to the real-time IoT-based day-to-day applications. Further, increased use of multimedia data in social networking, especially image data arises as a new source of big data. The storage, analysis, and retrieval of such data will be a challenge in the coming era (Table 1).

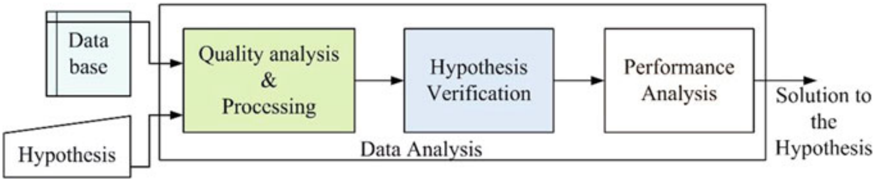


Fig. 3 Data analysis in big data Era

1.2 Data Processing in Big Data Regime

The problem of voluminous data processing has been solved by the data science in the last decade. The data science used various tools for the analytics of such a huge data. Figure 3 depicts the block diagram of a typical data processing architecture.

However, *value* of the gathered data will play a big role in next-generation communication technology. The principle issue in big data is its storage and processing. The data storage aspect is resolved by the advent of cloud-based distributive computing. The availability of the voluminous data from various sources poses a challenge of its category-wise processing and further analysis to decide its usability (value). To decide the value of the gathered data and to ripe maximum output from it, the notion of predictive and prescriptive analysis will be useful.

Section 2 describes the fundamentals of prescriptive and predictive analytics. The upcoming research challenges and the open issues in the prescriptive and predictive analytics are described in Sect. 3, and the paper is concluded in Sect. 4.

2 The Predictive and Prescriptive Analytics in Big Data

2.1 The Predictive Analytics

Data analytics has a wide range of methodologies and techniques to analyze the available data. They are broadly classified as descriptive, diagnostic, predictive, and prescriptive analytics [4]. Figure 3 shows the evolution of the predictive and prescriptive analytics.

The descriptive analytics is generally performed at the initial stage of data accumulation. It helps out to get a reasonable insight into the nature and pattern of the data. The descriptive analytics mainly concentrated on "*what*" with the help of classification, clustering, and segmentation of the data.

Once it is understood "*what*", the next stage is to concentrate on "why". This aspect of analytics is known as diagnostic analytics, which is carried out with the help of machine learning algorithms. Both the descriptive analytics and diagnostic analytics are largely dependent on the incidences, which have already occurred in the past. However, these two approaches have limitation to predict the futuristic behavior of the user and the cause-action strategy to deal with it. Here comes the need of prediction-based forecast and the redial action (prescription of the pattern) for the user behavior based on available data. Predictive analytics uses machine learning algorithms and statistical analysis techniques to analyze current and historical data to make predictions about future trends, behavior, and activity.

2.2 *The Prescriptive Analytics*

The business opportunities are very stringent, and hence the entrepreneurs always wish to capitalize the available opportunities. However, having the insight on previous behavioral patterns and futuristic trend forecast will not be sufficient to take advantage of business opportunities. There is a trade-off between the insights generated and optimal operational courses of actions [5]. Hence, it is the need of the hour to utilize the analytics to transform information into valuable data (insights) so as to act upon them effectively to meet their objectives [6–8]. Prescriptive analytics helps to overcome this gap and is considered as the next frontier in the business analytics [9].

Prescriptive analytics is related to both descriptive and predictive analytics. While descriptive analytics aims to provide insight into what has happened and predictive analytics helps to model and forecast what might happen, prescriptive analytics seeks to determine the best solution or outcome among various choices, given the known parameters.

It provides organizations with adaptive, automated, and time-reliant sequences of operational actions. It answers the questions like "What", "Why", and "when" and very importantly "how" should be done [10]. However, in the preview of big data, analytics engine must be dynamic enough to provide the best possible cause–effect relationship for the decision-makers in the enterprise. This requirement can be met by the virtue of an adaptive mechanism, which will best utilize the predictive and prescriptive analytics at a given time of instance. Figure 4 shows the conceptual block diagram of an adaptive prescriptive and predictive analytics mechanism (Table 2).

Table 2 Predictive versus prescriptive analytics

Parameter	Predictive analytics	Prescriptive analytics
Nature of outcome	Provides only a prediction of outcome	Provides deterministic outcome
Base data requirement	Historical and present data	Historical and present data
Type of insights	Projections of various outcomes	Ascertainment of a set of outcomes
Processing of big data	Machine learning algorithms	Artificial intelligence
Decision-making process based on	Descriptive analytics	Both descriptive and predictive analytics

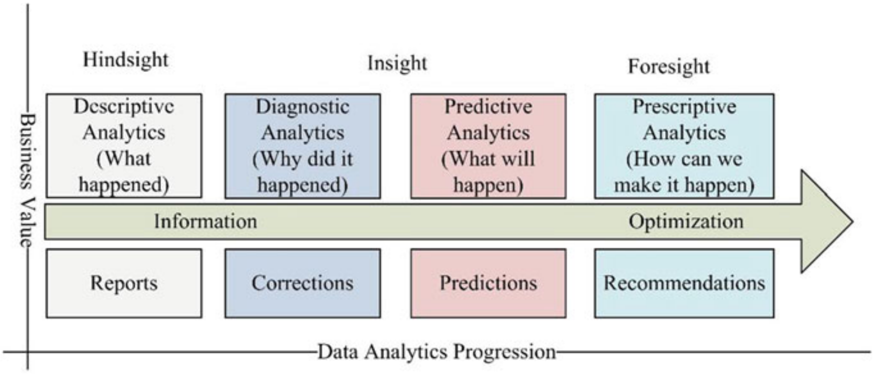


Fig. 4 The predictive and prescriptive analytics

2.3 The Prescriptive Analytics Platforms

There are several commercial and open-source platforms available for the predictive and prescriptive analytics. This section briefly reviews the various available tools and their features. A beginner may use one of these tools for the research work (Tables 3 and 4).

Table 3 Predictive analytics tools

Name	Technology	Key features	Availability
RapidMiner studio	Machine learning	Visual workflow design and unified platform	Proprietary
KNIME platform	Modular data pipelining	Churn analysis, credit scoring, and social media sentiment analysis	Open source
IBM predictive analytics	Statistical analysis	Hands-on predictive analytics and direct deployment into business processes	Proprietary
Minitab	Statistical analysis	Smart data import and automatic graph updating	Proprietary
TIBCO spotfire	Data mining	Location analysis, big data analysis, and data wrangling	Proprietary
DataRobot	Machine learning	Drag and drop dataset and machine learning automation	Proprietary

Table 4 Prescriptive analytics tools

Name	Technology	Key features	Availability
Ayata	Machine learning and operation research	Adapted and automated, integrated prediction and prescription, and secure SaaS delivery model	Proprietary
AIMMS	Web-based solutions	Customized saluting, on-site or cloud-based support, and virtual solution environment	Proprietary
NGData	Personalization and engine software	Adaptive and just in time operation	Proprietary
LIONoso	Machine learning	Healthcare solutions, R&D facility, learning from data, and optimization	Proprietary
Profitect	Data warehousing	Intelligent data transform, multiple business technology support, and real-time customer feedback	Proprietary
IBM prescriptive analytics	Data mining and machine learning	Automate complex decisions and trade-offs, risk analysis, increase agility, and efficient resource management	Proprietary

3 The Open Research Issues of Predictive and Prescriptive Analytics

Big data is available in almost all aspects of life as data generation takes place in every form of human behavior. It benefits both research and industrial fields such as health care, financial services, and commercial recommendations. Slowly but steadily, business and enterprise houses are adopting the prescriptive analytics as the best alternative to drive valuable insights from the user-driven big data. This paves the way to explore many new avenues for research and development and subsequent employability aspect too. The trusted area for the futuristic research will be as follows:

A. *Smart City*

The notion of smart cities has paved the ways for great changes in fundamental facilities of transportation. Possessing certain features such as a large scale, diversified foreseeability, timeliness, city traffic data, sanitation, and security represent the scope of big data. To ensure hassle-free life, accurate decisions need to be initiated in these areas, which in turn depend on the valuable insights available from the big data generated from the users.

B. *Healthcare Applications*

The notion of automated health assistance is now becoming prevalent all over the globe. With the advancement of IoT technologies, tiny body area sensors are deployed with the users, which continuously monitor the human vital signs. Mechanisms are under test, which generates automated alerts for the human health to the self as well as the medical service providers. In this process, a huge data is generated in terms of basic signals and images too. The storage, processing, and retrieval of such a huge data will be the next-generation research challenge. With the help of predictive and prescriptive analytics, pinpoint decisions may be initiated to provide better services to the users.

C. Content Recommendation

Increased use of multimedia content on social network has now acquired a day-to-day activity of human life. Users are now more cautious about expressing themselves. For this purpose, users seek the help of various online services for multimedia applications. With the help of predictive and prescriptive analytics, service providers may put forward the best alternatives for the users.

D. User Behavior Prediction

Many of the network big data predictions are based on data from online social networks. Big data is used for predictions based on ranked data such as elections, car performance, and other areas in business and politics. Using the predictive and prescriptive analytics, policy-makers may frame deceive policies regarding the various sociopolitical, economical, and educational strata.

E. Security and privacy of user data

With the global business digitalization driving the corporate world toward the risk of cyberattacks more than ever before. The predictive and prescriptive analytics of big data has the potential to offer protection against the cyberattacks. Based on the huge data, the security service providers can easily predict the behavior of eavesdroppers and prescribe the corrective measure against them. It is believed that the security analytics will emerge next generation.

F. Application of Artificial Intelligence

Along with big data, the notion of artificial intelligence (AI) and machine learning is storming the world these days. These are the set of technologies that empower connected machines and computers to learn, evolve, and improve upon their own learning by reiterating and consistently updating the data bank through recursive experiments and human intervention. These characteristics empower the machines to predict and thereon prescript the user trends and behavior (Fig. 5).

In the future, it is expected that the IoE and big data boundaries will vanish and it will become a synonym. In view of this, the scope of research in the area of

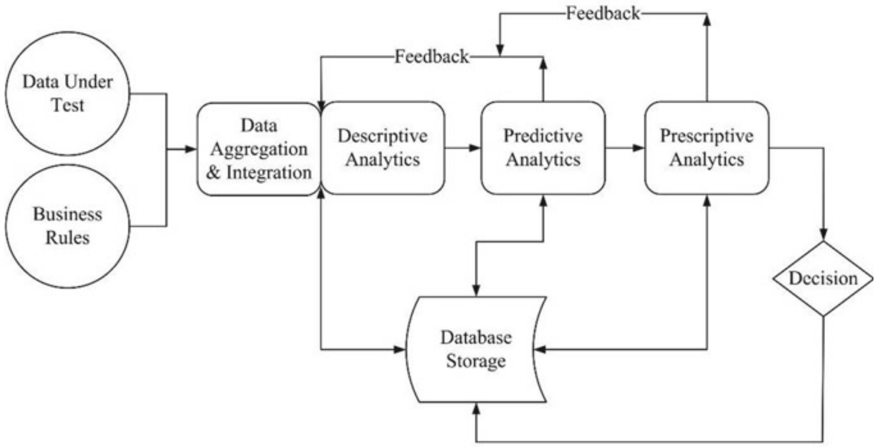


Fig. 5 Conceptual block diagram of a prescriptive analytics engine

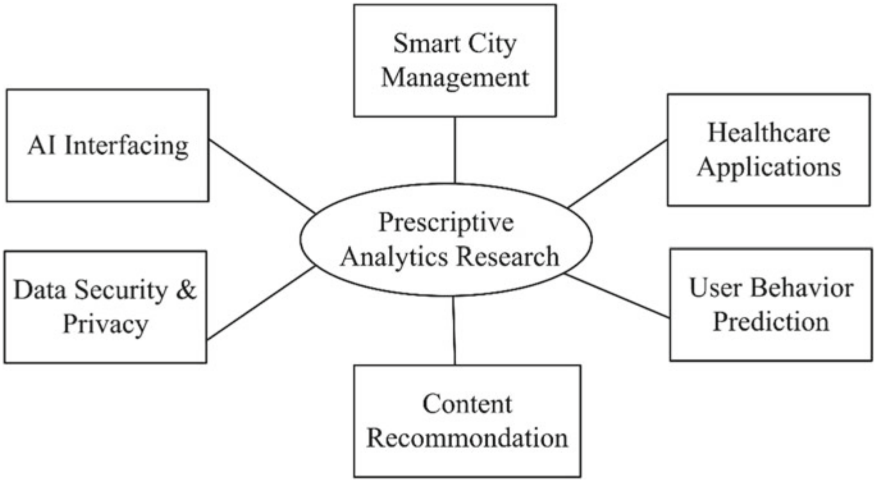


Fig. 6 Research directions in the prescriptive analytics

prescriptive analytics will accommodate all areas of engineering and technology, science, and humanities (Fig. 6).

4 Conclusions

The paper reports the predictive and prescriptive analytics in big data. It is believed that the next-generation data processing technology will be largely dominated by the predictive and prescriptive analytics along with big data and IoT. The concept of IoE will be totally dependent on the predictive and prescriptive analytics. In future, research may be undertaken on the AI-based prescriptive and predictive analysis of big data.

References

1. Deshpande, P., Iyer, B.: Research directions in the internet of every things (IoET). In: International Conference on Computing, Communication and Automation (ICCCA), Noida-India, pp. 1353–1357, May 2017
2. Number of mobile phone users worldwide 2013–2019. https://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide
3. Cox, M., Ellsworth, D.: Application Controlled Demand Paging for Out-of-Core Visualization. NASA Ames Research Centre Report, Mountain View, CA (1997)
4. Hardoon, D., Shmueli, G.: Getting Started with Business Analytics: Insightful Decision Marking. CRC Press, Boca Raton, FL (2013)
5. Barga, R., Fontama, V., Tok, W.H.: Predictive Analytics with Microsoft Azure Machine Learning: Build and Deploy Actionable Solutions in Minutes. Apress (2014). ISBN-13 (pbk): 978-1-4842-1201-1 and ISBN-13 (electronic): 978-1-4842-1200-4
6. Baker, P., Gourley, B.: Data Divination: Big Data Strategies. Delmar Learning (2014). ISBN: 1305115082 9781305115088
7. Chen, H., Chiang, R.H., Storey, V.C.: Business intelligence and analytics: from big data to big impact. MIS Q. **36**(4), 1165–1188 (2012)
8. Kaisler, S.H., Espinosa, J.A., Armour, F., Money, W.H.: Advanced analytics-issues and challenges in a global environment. In: 2014 47th Hawaii International Conference on System Sciences (HICSS), pp. 729–738. IEEE (2014)
9. Evans, J.R., Lindner, C.H.: Business analytics: the next frontier for decision sciences. Decis. Line **43**(2), 4–6 (2012)
10. Delen, D., Demirkan, H.: Data, information and analytics as services. Decisi. Support Syst. **55**(1), 359–363 (2013)

Indexing in Big Data

Madhu M. Nashipudimath and Subhash K. Shinde

Abstract Nowadays communication is through social media for almost all activities like business, knowledge, personal updates, etc. This leads to the generation of large amount of data related to different activities. Hence, social media have become a vital content of our life. But going through this huge data for analysis is a tedious and complex task. There are many solutions to overcome this problem. Data reduction, indexing, and sorting can be the solutions. Further, which will be used for visualization, recommendation, etc. Indexing techniques for highly repetitive data group have become a relevant discussion. These techniques are used to accelerate queries with value and dimension subsetting conditions. There are different types of indexing with the suitability of data type, data size, dimension, representation, storage, etc. Indexing is of vital need as whatever electronic text collection is available, it is mostly large scale and heterogeneous. Hence, the motto is to find an improved approach for text search as it is used right from the help services built into operating systems to locate file on computers. Tree-based indexing, multidimensional indexing, hashing, etc., are few indexing approaches used depending on the data structures and big data analysis (BDA). Indexing's need is to address the speed of search. So, size of index shall be a fraction of original data and to be built at the speed of data generation to avoid delay in result. Here, few indexing techniques/search structures are discussed based on data structure, frame work, space need, simplified implementations, and applications.

Keywords Algorithms · Big data analysis · Data indexing · Data structure Querying

M. M. Nashipudimath (✉)
Faculty of Computer Engineering, Pacific Academy of Higher Education and Research University, Udaipur, India
e-mail: madhu.nashipudi@yahoo.in

S. K. Shinde
Computer Engineering, Lokmanya Tilak College of Engineering, Navi Mumbai, India
e-mail: skshinde@rediffmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_15

1 Introduction

Collection of belief and variant is related in some meaning and differ in some other and can be defined as data. Collection and analysis are carried out for decision-making. So, data is used to provide a wealthy asset for the knowledge finding and decision-making. A database is well organized to make access, manage, and update task easily. Big Data (BD) is a fresh grammatical element used to refer the datasets that are larger in size like yottabytes (YB) and have better complexity [1]. As we are in the BD age, there is a nascent need for knowledge extraction from huge data.

The present methods of BD process can be applied with some improvements to overcome some characteristics of the issues as shown in Fig. 1 [2].

As mentioned in Fig. 1, indexing plays an important role in order to speedup the execution time. Indexing is preferred as an efficient method to overcome the cost of disk operations like reading, writing, deletion, insertion, modification, query speeds, etc. It can be used in both traditional relational databases that are used to manage structured data and other technologies that manage semi-structured and unstructured data. In Big Data, indexes are a list of tags, names, subjects, etc., of a dataset with references, where data can be found. The purpose of Big Data indexing is to share the datasets according to a measure that will be used often in query.

Cost may be added for maintaining and updating index file dynamically if required. Data is stored in a distributed system when a single machine finds difficulty to hold the BD. So, the index of BD is supposed to be built on distributed system. Similarly, new query theory should be bucked up because BD not only means huge in size but it is also complex in structure, high dimensional, and heterogamous. Hence, it is no more advisable to have longstanding techniques for indexing with small structured datasets. To face these challenges, researchers have come up with several methods depending on priorities like dimension, structure, time, data complexity, space, etc.

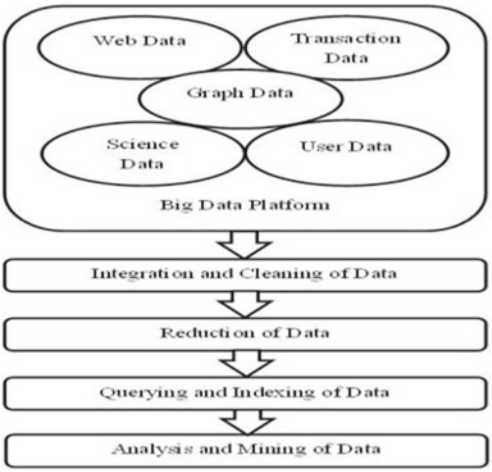


Fig. 1 Big data process

2 Indexing

Dimension setting, space saving, and simple implementation-based queries are used to extract the information for BDA. But it brings different challenges in different levels of data analysis due to multimedia data size, and the duplicate and similarity measures in data are extracted [3, 4].

Queries with value and dimension subsetting conditions are commonly used by researchers to find the useful information from BD. Indexing methods play an important role in supporting queries on high-dimensional BD efficiently. Further, some indexing is applied in BD to carry extraction tasks from huge complex scalable datasets with capability, and hence it is merely difficult to perform manual exploration on such data. Prompt high-throughput indexing technique would affect the performance of data query results [5]. Hence, a suitable indexing technique is essential to effectively access huge data such as semantic indexing-based approach [6], bitmap indexes [7] graph query processing [8], etc.

There are numerous indexing techniques available. Still, research is going on for progressive survey that finds the action and outcome of techniques to handle indexing in BD issues.

3 Indexing Techniques

The recent work in indexing suggests the way to improve the performance without comprising in quality. Few of such techniques are discussed further along with the dataset used.

A method, which relies on the selection followed by the order of a set of reference objects by their distance to every database objects named as permutation-based indexing [9]. These ordered permutations are used to understand the computation of the exact ordering distance values in between the submitted query and the objects related to database. The number of reference points plays an important factor in permutation-based indexing. These points have to be increased with respect to data size. Similarly, aggregation and summary of the ordered lists per object are also required, specifically for large-scale data.

The quantizing scheme for the data into a given number of buckets in order to summarize concisely and use is still to be investigated. Approximation is made with the help of ranking of the values of distance between the objects and reference objects, which is well exposed by Mohamed et al. [10]. A candidate selection is carried out as a result for a K-NN query with the help of data structure called metric permutation table (MPT). Work exhibits that the MPT performs better with respect to search time even if the number of reference objects increases; however, it requires more memory. Hence, the performance of this technique, reference points, and its distribution are interdependent in the dataset and the indexing technique. In order to reduce the search time, this work can be experimented with clustering or any other unsupervised approaches.

String B-tree [11] acts similar to combo of B-trees and Patricia trees for internal-node indices, where addition of extra pointers is done to speedup the search and update the operations. This is successfully applied to database indexing.

Adaptive semi-supervised recursive tree partitioning (ART) framework [12] is recommended for BD set like patients' information. This helps to identify and acquire patients with identical clinical or diagnostic structure. The framework is configured for semi-supervised settings with better performance with respect to time and retrieval. Tree-based methods are mostly unsupervised, follow euclidean space, and do not consider label information. But this is not the case in tree-based approach recommended by authors [13].

The tree is built recursively with following two terms to make it more efficient. (1) Supervised term for indexing, i.e., prior supervision knowledge (2) unsupervised term, i.e., geometric structure of the patient vectors. Binary space partitioning tree is built with the above two terms along with a constant λ to balance the contributions from both terms. It is a common problem in partition-based method that the deeper indexing tree is created by high dimensions. It takes long retrieval time. So, the sampling-based approach is used to overcome the problem of huge data vector. Further, k-means clustering is added with data matrix with the reduced data dimensionality. Eventually, this leads to the efficient time management. This framework can also handle huge data until it is represented in terms of vectors.

Data management [14] of video using spatial indexing and field-of-view (FOV) query is carried out, where video frame is labeled with spatial extent of its reporting area termed as field-of-view (FOV). These video frames are trimmed into slices of pie with location and orientation information. Hence, conventional spatial indexing approach such as R-tree is not suitable for indexing efficiently. For example, multi-level grid-based indexes can take care of location and orientation, but cannot manage skewed distribution because the distribution of UGVs' locations is nonuniform as it is used in popular locations. Hence, the authors worked for advanced R-tree to tackle locations, orientations, and view distances in rotation to filter and optimize with better scalability and efficiency.

The proposed R-tree focuses on the location points identified by camera that are added as index nodes to store the orientations. The R-tree generates small minimum bounding rectangle (MBR) to minimize their dead spaces while performing the orientation filtering.

Optimization technique adds a second variable that uses orientation information, which is achieved during node split and merge operations. FOVs' viewable distances are added as third variable during the filtering and optimization process.

In augmenting R-tree, position length and extra positioning are the basic needs in the enhancement of the index performance. Optimization techniques are used to demonstrate this fact in augmenting R-tree. The pruning of an object/index node and FOV object/index node report (all the FOV objects in the index node) are the possible solution(s). This prune or report decision is taken by algorithm. Pruning activity is carried out, if FOV object overlaps with the range query object. R-tree index node is decided by pruning and total hit strategies. Utilization of orientation and view distance information are key activities for the merge and split operations. Hence, the second and third variations significantly improve the demonstration of the recommended approach compared to R-tree and grid.

Regression [15] determines the index and weight in the index compilation method of BD. Authors have also worked on analysis of multiple integration strategies for index compilation. Strategies like integration of BD and classical methods, data integration, integration of index system, and integration of data weight are analyzed.

R. Suganya Devi et al. have proposed a prompt method to crawl and index [16] the links associated with the specified URLs. Depth-first search is used to crawl the links from the specific URL that leads to hierarchical scanning of web links. Metadata such as title, keywords, and description are extracted and accessed via source code for the analysis work to be carried out as a result of web crawling. Building the database of web pages and links of WWW are carried out followed by recrawling to keep the contents of the database current. One can put effort to minimize the bandwidth required and make it accessible to the next level of links.

Sparsifying dictionary [17] is used for inverted indexing, where particular regression [18] is applied to allocate the text files to buckets. The optimized dictionary is applied to decrease the data dimensionality. Use of sparsifying dictionary is carried out by focusing on the strength of the original vector on a few coefficients of a high-dimensional state. Results have better quality and improved computational time compared to the well-known indexing trees.

The iMinMax(θ) [19] is a tunable index schema. Points are found in high-dimensional spaces to map with the points in single dimension values. Maximum or minimum values among all dimensions are determined to find the points in high dimensions based on the edges. Stable θ for each dimension is suggested for better result.

Value and dimension subsetting queries play an important role in mining information from multidimensional array. Aim is to accelerate the query with value and dimension subsetting conditions. A 2D-Bin will build an index for the blocks' value ranges, which is an efficient way to avoid accessing unnecessary blocks for value subsetting queries.

Swift array [20] divides multidimensional array into blocks, where each block stores the sorted values as it will reduce the cost of processing value subsetting queries. It is used as storage layout with indexing, which will help to speed the query execution. Hilbert space-filling curve [21, 22] is applied to improve the data locality for dimension subsetting queries. As a future development, compression techniques can be used as support to Swift array in order to reduce the data in the block.

Skyline algorithm with sorted positional index lists (SPIL) [23] is used with reward of index-based algorithms and generic algorithms. It acquires interesting points from an immense data positioning. SPIL is fetched by skyline criteria until the candidate positional index is found in all of the participating lists. The SPIL is generated for each attribute and is arranged in ascending order of attribute. Retrieval is carried out in round-robin fashion. Skyline will eliminate the candidate positional index whose corresponding tuple is not a skyline result, as SSPL will figure out the scan depth of the involved SPIL.

In the succeeding level, SPIL will perform an ordered and selective scan on the table. This action will refer the candidate positional indexes, which is obtained in previous level to compute skyline results. As compared to LESS [24], skyline needs added activity called preprocessing although it performs better on BD as compared to LESS. Improvement can be worked on preprocessing with reduced implementation complexity and space requirement. Multilevel key ranges are referred by fast indexing technique. Multiple searches can be performed by using single pass on the indexing structure to reduce the price of execution.

This method works for any large number of objects' sorting based on keys. Implementation is based on the B + -tree technique, which is similar to indexed sequential access method (ISAM).

An approach of FB + -tree [25] is quite competent for performing regular searches without going to disk. Hence, it suits well for high buffer size, appropriate tree configuration, and popular queries.

Multimedia exploration framework [26] is suggested for social media analysis in emergency management with online analysis (e.g., by standard incremental term frequency–inverse document frequency), geo-tagging by entity recognition related to a location, and an interactive visualization of the results. Events are clustered by location and followed by online sub-event detection/new terms, so that the rarely used terms can be removed and new essential terms are added. This activity is followed by dynamic indexing with high number of indexing terms followed by learn and forget method as a suggestive idea. In future, this art can be automated by reorientation of the parameters for learn and forget task.

Ideas of phrase indexing method [27] based on the term discrimination model are discussed by Fagan and Joel, where non-syntactic phrase construction method and related issues are covered. Whereas, Al-Shalabi et al. [28] have discussed 3- and 4-g term indexing. The authors have presented an innovative idea of single term and dual term query to calculate the similarity between query and documents. Martin Krallinger et al. [29] have provided an extensive and clear description of fundamental concepts, technical execution, and current technologies for targeted retrieval of chemical documents through automatic recognition of chemical entities in the text. Few selective articles are represented in Table 1 as sample review.

4 Discussion

Data accumulation and handling is a present-day challenge and indexing may be one of the solutions for improved search. Due to heterogeneity and size, the indexing plays an important role to reducing the search time, writing, memory usage, etc. Prior to indexing, sorting out the noise and validating the veracity of the data are issues to be sorted out. Ranking based on relevance and identification of tricks in unstructured data are areas to work. Scope exists in schema mappings [10] to work due to the

Table 1 Sample review

Literature	Features	Advantages	Constraints/future scope
Ferragina et al. [11]	B-trees and Patricia trees are combined for internal-node indices	Efficient with parallel-disk model	Updating may be time consuming, whereas search is fast
Han et al. [23]	Skyline criteria till candidate positional index are found in all of the participating lists. Retrieval in round-robin fashion	Low-space overhead to decrease the I/O cost importantly	Preprocessing to minimize complexity and space requirement
Geng et al. [3]	Swift array is used and blocks are placed in the order of a Hilbert space-filling curve to improve data locality	It is suitable for large datasets	compression and MPI (message passing interface)-IO can reduce the data in the block
Mohamed and Marchand-Maillet [10]	Quantization schema is used. The technique is associated with MPT as a simple data structure for fast response time and effective memory usage	Minimum memory usage and improved search precision	Clustering or any unsupervised approaches may reduce search time
Devi et al. [16]	Crawling the links and retrieving all information associated with them is carried to facilitate easy processing for other uses	DFS algorithm is used for complete hierarchical scanning of corresponding web links	Minimize the bandwidth required and make it accessible to the next level of links
Wang [12]	The tree structure is recursively built with two terms (1) supervised term for indexing (2) unsupervised term from patient vector	Framework is configured for semi-supervised settings with better performance	Time consuming to follow indexing with prior supervised knowledge regarding pairwise constraints
Lu [14]	Spatial indexing and querying of field-of-view (FOV). frames are trimmed into slices of pie with location and orientation	Advanced R-tree is used to tackle locations, orientations, and view distances with scalability and efficiency	A method for flock insertion of video can be experimented to index huge videos
Borge [18]	Large-scale high-dimensional indexing algorithm based on sparse approximation and inverted indexing	Dictionary is optimized to reduce dimensionality	Addition of learned or customized dictionary for indexing
Pohl [26]	Clustered by location and online sub-event detection/new terms to add new terms and vice versa	Dynamic indexing with high number of indexing terms followed by learn and forget method	Automate by reorientation of the parameters for learn and forget task

limitations of artificial intelligence, which are generated by automatic processes. These can be worked out by domain experts. Block size plays an important role in dimension subsetting queries [3]. Hybrid tree approach [11] utilizes more memory, and hence limits to minimum dimension access. Inverted indexing stands as support for high dimensions but takes more time to process and limits the search space. Redrawing the changes [16] in web page consumes time particularly during flock insertion.

Many repositories provide data in different levels to test the methodology, so the selection of right repositories is essential. New schema [19] are introduced to map the points from high dimensions to single dimensions which need stable θ. Sorted positional index lists [23] need an activity called preprocessing with reduced implementation complexity and space requirements. Learned and customized dictionary can support indexing apart from reducing the dimensions [18]. Bandwidth with proper scale is to be provided to handle the redraw changes [16] in heterogeneous data. Techniques are to be addressed for flock insertion of data [14] during online process. Handling semantic conflicts [30, 31] using LSI can be extended with taking care of scalability and performance to support semantic retrieval of documents.

5 Conclusion

This discussion is focused on frameworks, data structures, types of arrays, tree structures, and mapping approaches for indexing. With the knowledge of automatic phrase indexing [32] and other discussions [31, 33], a content is in thought process to implement a recommendation system, where few constraints of discussed methods can be overcome.

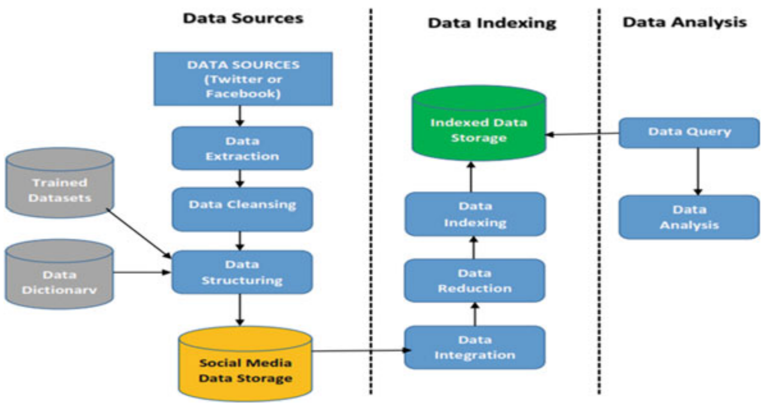


Fig. 2 Integration and indexing model

As mentioned in Fig. 2, we target to design a semantic data relations learning (SDRL) module to learn the semantic data relation among the social data to efficiently classify data for accurate integration and indexing. The parameter-based refined set of social data is then matched against a semantic relation data knowledge as a part of the semantic similarity-based indexing. A key part of this process involves enhancing in the data indexing. The basic idea of the proposed approach is to improve the indexing with semantic relevant data terms, and then find the similarity measure of the semantically enhanced indexing with the classified social data description vectors generated in the semantic annotation refinement phase. For evaluating this similarity, we will employ LSI [31]-based Indexing technique that uses cosine measure as the similarity metric.

References

1. Mining Big data in Real time: Albert bifet. Informatica **37**, 15–20 (2013)
2. Lodha, R., Jain, H., Kurup, L.: Big data challenges: data analysis perspective. Int. J. Curr. Eng. Technol. **4**(5), 3286–3289 (2014)
3. Deshpande, P., Sharma, S.C., Peddoju, S.K.: Efficient multimedia data storage in cloud environment. Inform. Int. J. Comput. Inf. **39**(4), 431–442 (2015)
4. Deshpande, P., Sharma, S.C., Peddoju, S.K.: Data storage security in cloud paradigm. In: Proceedings of Fifth International Conference on Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing, vol. 436. Springer, Singapore (2016)
5. Wang, M., Holub, V., Murphy, J., O'Sullivan, P.: High volumes of event stream indexing and efficient multi-keyword searching for cloud monitoring. Future Gener. Comput. Syst. **29**(8), 1943–1962 (2013)
6. Rodríguez-García, M.Á., Valencia-García, R., García-Sánchez, F., Samper-Zapater, J.J.: Creating a semantically-enhanced cloud services environment through ontology evolution. Future Gener. Comput. Syst. **32**, 295–306 (2013). https://doi.org/10.1016/j.future.2013.08.003
7. Wu, K., Shoshani, A., Stockinger, K.: Analyses of multi-level and multi-component compressed bitmap indexes. ACM Trans. Database Syst. **35**(1), 1–52 (2010). https://doi.org/10.1145/1670243.1670245
8. Cheng, J., Ke, Y., Fu, A.W.-C., Yu, J.X.: Fast graph query processing with a low-cost index. VLDB J. **20**(4), 521–539 (2011)
9. Gonzalez, E., Figueroa, K.: G. Navarro.: Effective proximity retrieval by ordering permutations. IEEE Trans. Pattern Anal. Mach. Intell. **30**(9), 1647–1658 (2008). https://doi.org/10.1109/TPAMI.2007.70815
10. Mohamed, H., Marchand-Maillet, S.: Quantized ranking for permutation-based indexing. Inf. Syst. (Elsevier) **52**, 163–175 (2015)
11. Ferragina, P., Roberto, G.: The string B-tree: a new data structure for string search in external memory and its applications. J. ACM (JACM) **46**(2), 236–280 (1999)
12. Wang, F.: Adaptive semi-supervised recursive tree partitioning: the ART towards large scale patient indexing in personalized healthcare. J. Biomed. Inf. **55**, 41–54 (2015)
13. Bentley, J.: Multidimensional binary search trees used for associative searching. Commun. ACM **18**(9), 517 (1975)
14. Lu, Y., Shahabi, C., Kim, S.H.: Efficient indexing and retrieval of large-scale geo-tagged video databases. GeoInformatica **20**(4), 829–857 (2016)
15. Mei, Y., Jing, Z., Liu, J.: Research on index compilation method of integrating big data. Manag. Eng. **22**, 7 (2016)

16. Devi, R.S., Manjula, D., Siddharth, R.K.: An efficient approach for web indexing of big data through hyperlinks in web crawling. Sci. World J. (2015)
17. Borges, P., Mourao, A., Magalhaes, J.: High-dimensional indexing by sparse approximation. In: Proceedings of the ACM ICMR, pp. 163–170. ACM (2015)
18. Borges, P., Mourão, A., Magalhães, J.: Large-scale high-dimensional indexing by sparse hashing with l0 approximation. Multimed. Tools Appl. 1–24 (2016)
19. Ooi, B.C., Tan, K.L., Yu, C., Bressan, S.: Indexing the edges—a simple and yet efficient approach to high-dimensional indexing. In: Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 166–174. ACM (2000)
20. Geng, Y., Huang, X., Yang, G.: Swiftarray: accelerating queries on multidimensional arrays. Tsinghua Sci. Technol. **19**(5), 521–530 (2014)
21. Moon, B., Jagadish, H.V., Faloutsos, C., Saltz, J.H.: Analysis of the clustering properties of the Hilbert space filling curve. IEEE Trans. Knowl. Data Eng. **13**(1), 124–141 (2001)
22. Lawder, J.K.: Calculation of mappings between one and n-dimensional values using the hilbert space-filling curve, Technical report JL1/00 Birkbeck College, University of London (2000)
23. Han, X., et al.: Efficient skyline computation on big data. IEEE Trans. Knowl. Data Eng. **25**(11), 2521–2535 (2013)
24. Godfrey, P., Shipley, R., Gryz, J.: Algorithms and analyses for maximal vector computation. VLDB J. **16**(1), 5–28 (2007)
25. Yu, C., Boyd, J.: FB + -tree: indexing based on key ranges. In: IEEE 11th International Conference on Networking, Sensing and Control (ICNSC). IEEE (2014)
26. Pohl, D., Bouchachia, A., Hellwagner, H.: Online indexing and clustering of social media data for emergency management. Neurocomputing **172**, 168–179 (2016)
27. Fagan, J.L.: Automatic P hrase indexing for document retrieval: an examination of syntactic and non-syntactic methods. In: ACM SIGIR Forum, vol. 51, no. 2. ACM (2017)
28. Al-Shalabi, E.F.: The effect of N-gram indexing on Arabic documents retrieval. Int. J. Comput. Sci. Issues (IJCSI) **14**(1), 115 (2017)
29. Krallinger, M., et al.: Information retrieval and text mining technologies for chemistry. Chem. Rev. (2017)
30. Adamu, F.B., Habbal, A., Hassan, S., Les Cottrell, R., White, B., Abdullahi, I.: A survey on big data indexing strategies. In: 4th International Conference on Internet Applications, Protocol and Services (NETAPPS2015), Cyberjaya, Malaysia (2015)
31. Paliwal, A.V., Adam, N., Bornhoevd, C.: Adding semantics through service request expansion and latent semantic indexing. In: Proceedings of IEEE International Conference Services Computing (SCC), July 2007
32. Gani, A., et al.: A survey on indexing techniques for big data: taxonomy and performance evaluation. Knowl. Inf. Syst. **46**(2), 241–284 (2016)
33. Jagadish, H.V., Mendelzon, A.O., Milo, T.: Similarity-based queries. In: Proceedings of the Fourteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, PODS'95. ACM, New York, USA (1995)

DataSpeak: Data Extraction, Aggregation, and Classification Using Big Data Novel Algorithm

Venkatesh Gauri Shankar, Bali Devi and Sumit Srivastava

Abstract A huge amount of data is coming due to large set of computing devices. As a birth of the variety of data, data processing and analysis is a big issue in big data analytics. On other hand, data consistency and scalability is also a major problem in the large set of data. Our research and proposed algorithm aims to data extraction, aggregation, and classification based on novel approach as "DataSpeak". We have used k-Nearest Neighbors with Spark as reference and produced a novel approach with modified algorithm. We have analyzed our approach on the large dataset from travel and tourism, placement papers, movies and historical, smartphone, etc., domains. As for ability and accuracy of our algorithm, we have used cross validation, precision, recall, and comparative statistical analysis with the existing algorithm. Our approach returns with the fast accessing of data with efficient data extraction in a minimal time when compared to the existing algorithm in same domain. As concerned with the data aggregation and classification, our approach returns 98% of data aggregation and classification based on the data structure.

Keywords Big data · Big data analytics · Classification · kNN
Spark framework

1 Introduction

Data is characterized as the amounts, characters, or images on which operations are performed by a PC, which might be put away and transmitted as electrical flags and recorded on attractive, optical, or mechanical account media as Google show us [1].

V. G. Shankar (✉) · S. Srivastava
SCIT, Manipal University Jaipur, Jaipur, India
e-mail: venkateshgaurishankar@gmail.com

B. Devi
CSE, Jayoti Vidyapeeth Women's University, Jaipur, India
e-mail: baligupta03@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_16

The idea of big data is not much; as the name proposes, “Big Data” alludes to overflowing measures of data which are too extensive to ever be handled and investigated by customary devices, and the data is not put away or oversaw effectively. Since the measure of big data increments exponentially, more than 500 TB of data is transferred to Facebook alone, in a solitary day, it speaks to a genuine issue as far as investigation [2, 3].

1.1 Classification of Big Data

Classification is important description for measurement of varieties. Nowadays, a large set of data is created from various online networking sites like Twitter, Facebook, and from various exchanges done in organization’s databases, from inventory network situations which supplies huge amounts of data, for instance, given a number of scanners and so forth. This huge amount of data as big data is also classified into the five types of big data. The first classification of big data is Structured data, concerns all data which can be put away in database with row and column. They contain relational key, and can be effectively mapped into pre-outlined fields. Structured data is exceptionally sorted out data that transfers conveniently into a relational database. It is moderately easy to insert, store, query, and examine, however, it must be entirely characterized as far as field type and name [4].

The next classification is Unstructured, which is the content written in different structures like web pages, messages, email messages, pdf records, word reports, etc. Unstructured data is likewise characterized in view of its source, into machine-created or human-produced. Machine-created information represents all the satellite pictures, the scientific data from different examinations, and radar data caught by different aspects of technology [4].

The third classification is Semi-Structured, which is a type of structured data that does not acclimate with the formal syntax of data models related to relational databases or different types of data tables, however, regardless contains labels or different markers to isolate semantic components and uphold pecking orders of records and fields inside the data [4].

The fourth one is Quasi-Structured, indicates the bunch of text data with volatile data style, which will be manipulated with work, framework, and interval. As for the example of Quasi-Structured data, the best example is website click-stream dataset, which will contain some impurities in dataset and structure [4].

At last, Quasi-Unstructured data is the eventual fate of big data, and is neither organized nor unstructured. Big data will be organized by instinctive strategies (i.e., “genetic algorithm”), or utilizing inherent patterns that rise up out of the data itself and not from rules forced on dataset by people [4]. See Fig. 1.

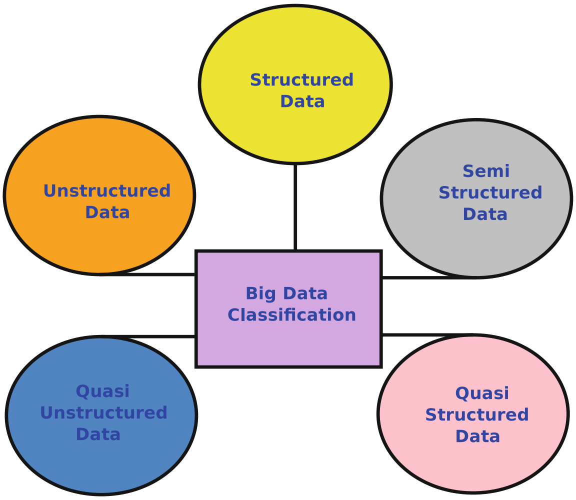


Fig. 1 Big data classification

1.2 Characteristics of Big Data

Volume [5] is presumably the best-known property of big data; this is nothing unexpected, considering more than 90% of all present data was made in the recent years. The present measure of information can really be very stunning. For example, as approximately 350 h video are uploaded to YouTube every minute. **Velocity** alludes to the speed at which data is being produced, delivered, made, or invigorated. Without a doubt, it sounds amazing that Facebook's data stores upwards of 300 petabytes of data, yet the speed at which new information is made ought to be considered. Facebook claims 600 terabytes of approaching data everyday. **Variety** [6] with regards to big data, we do not just need to deal with organized data, yet in addition, Semi-Structured and for the most part unstructured data also. Most of big data is by all accounts unstructured, yet next to sound, picture, video documents, online networking interface, and other text artifacts, there are additionally log records, click information, machine and sensor information, and so on. **Variability** [7], in case, of big data alludes to a couple of various things. One is the quantity of irregularities in the data. These should be found by peculiarity and anomaly detection techniques, if any important analysis happens. Big data is likewise factor in light of the large number of data measurements coming about because of numerous unique data sorts and sources. Variability can likewise allude to the conflicting pace at which big data is stacked into your database. **Veracity** is one of the awful qualities of big data. As any or the greater part of the above properties increment, the veracity (certainty or trust in the data) drops. This is like, yet not the same as, legitimacy or instability. Veracity alludes more to the provenance or dependability of the data source, its specific circumstance, and that it is so important to the analysis centered on it. Like veracity, **validity** alludes to how exact and rectify the data is for its intended usage.

As indicated by Forbes, an expected 60% of data scientist researcher's opportunity is spent purging their data before having the capacity to do any analysis. The advantage from huge data analytics is just tantamount to its hidden data, so you have to receive great data administration practices to guarantee predictable metadata data quality and normal definitions. **Vulnerability** [8, 9] in big data brings new security concerns. All things considered, a data rupture with big data is a major break. Tragically, there have been numerous big data ruptures or breaches like data theft, smartphone sensitive data leakage, and network data leakage. Because of the velocity and volume of big data, in any case, its **volatility** should be deliberately considered. You now need to set up rules for data currency and accessibility and also guarantee fast recovery of data when required. Ensure these are plainly attached to your business needs and procedures. With big data the expenses and complexity of a storage system and recovery process are amplified. Current big data **visualization** [5] tools and technology confront specialized difficulties because of restrictions of in-memory technology and poor adaptability, usefulness, and response time. You cannot depend on traditional graphs when endeavoring to plot a billion data of lines, so you require distinctive methods for representing information, for example, data classification and clustering or utilizing tree maps, sunbursts, parallel points, and coordinates, network or modeling diagrams. **Value** [5] is the end amusement for big data. In the reference of tending to volume, velocity, variety, variability, veracity, vulnerability, validity, volatility, and visualization, which takes a huge time, exertion, and assets—you need to make it certain that your business is getting an incentive from the data, i.e., noted as value of data.

2 State-of-the-Art: K-Nearest Neighbors

KNN can be utilized for both classification and regression predictive issues. In any case, it is all the more generally utilized as a part of classification issues in the business. KNN is very impressive in case of following three vital stages:

- Easily interpret output.
- Calculation time or execution time.
- Predictive capability.

KNN is based on nonparametric methods that are great when you have a tons of data and no earlier learning, and when you would prefer not to stress excessively over picking only the correct features [10, 11]. K selection is shown in Fig. 2.

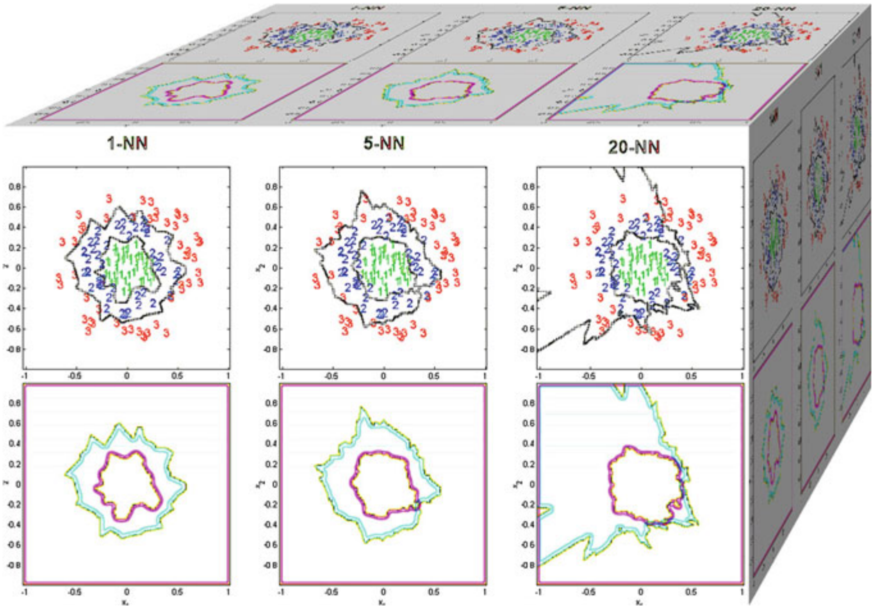


Fig. 2 KNN K-value selection

2.1 K-Nearest Neighbors: Working Steps

- Choosing the estimation of K in K-nearest neighbors is the most basic issue. A little estimation of K implies that noise will impact the outcome, i.e., the likelihood of over fitting is high. Consider the value of $k = n^{1/2}$.
- Arrange the calculated Euclidean distances in nondecreasing order for all training samples $\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$.
  - Sort the Euclidean distances and find the nearest neighbors with the set on minimum distance (value K).
  - Collect the family of all the nearest neighbors.
    Let k_i represent the number of points exist to the class among K points, i.e., $K \geq 0$.
  - Find the predicted score of the query object, i.e., the majority category of nearest neighbors (if $k_i > k_j$, i not equal to j), then value in class (i) [12].

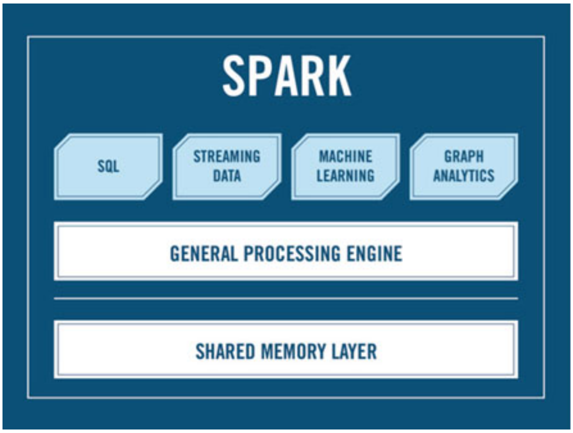


Fig. 3 Spark framework architecture

2.2 *K-Nearest Neighbors: Limitations*

- As soon as training set is very huge, it takes too much time in analysis [11].
- In case of test set data, the distance should be calculated between test data and all the training set data. Hence, so much time may be needed for the testing [12].

3 State-of-the-Art: Spark Framework

Apache Spark [13] is an open source, wide range data handling tool with uncovering advancement APIs that qualifies data specialists to fulfill extraction, machine learning, or SQL workloads, which request rehashed access to data collections. It is planned such that it can perform batch handling (preparing of the already gathered activity in a solitary cluster) and data streaming (manage data processing). It is a broadly useful, cluster processing tool.

Spark is planned such that it coordinates with all the big data tools. For instance, Spark can get to any Hadoop data source and can keep running on Hadoop clusters. Spark stretches out Hadoop MapReduce to the next level, which incorporates iterative enquiries and stream handling. Spark framework is shown in Fig. 3.

4 Related Work

Elegendy et al. [14] presented the features of big data in all perspectives with its generation and characteristic of big data in different scenarios. Demetrious et al. [15] proposed the article on big data with its opportunities in big data. They have explained many issues and good practices of big data in many contexts. He et al.

[16] gave his view on mobile network big data with the possible user social data, daily life mobile activity, and different app repository. EMC [17] showed a view for data science perspective and its option in future. It also explained many data science tools on big data analytics. Venkatesh et al. [18] proposed a framework as Anti-Hijack, which identifies session and intent vulnerabilities using honey pot technology. They analyzed big data of Android malware for security concern. Fu et al. [19] suggested a framework as EFANNA which is described as fast nearest neighbor-based algorithm, which emphasizes in divide and conquer techniques. They also took a large set of dataset as large data and the fast processing of big data with high throughput. Georgios et al. [20] demonstrated big data and its classification in the form of storage system and visualization. The state-of-the-art in this article is totally concentrated on cloud computing issues and options in social network as well as many data repository. Raghavendra et al. [21] addressed many big data aspects in computing with fast and efficient approach. The main scenario of this article is based on validation visualization, verification, and big data management. Yang et al. [22] presented the idea of new cloud storage challenges on big data platform. Innovation and challenges of big data consist of processing capability and analogical skills. This paper also emphasizes on digital earth sources with big data computing. Tsai et al. [23] explained about big data analytics survey and analysis in many scenarios. This paper also concentrated on many of data mining algorithms like k-means, kNN, etc. This paper also discussed many issues with many data mining algorithms.

Our novel approach is based on enhanced K-Nearest Neighbors algorithm (KNN). We have worked on many large sets of big data. We have proposed a fast extraction, aggregation, and classifiaction algorithm over KNN with Spark. In the next section, we will present the novel algorithms with high-performance result over three characteristics velocity, value, and volume.

5 Experimental Setup and Algorithm

Our proposed algorithm based on data extraction, aggregation, and classification based on novel approach as DataSpeak. Our approach has used K-Nearest Neighbors and producing a modified novel approach with Spark (which is taken as a retrieval tool). We have executed and analyzed our proposed framework on the large dataset from travel and tourism, placement papers, movies and historical, smartphone, geographical, satellite, genetic, click streamed domain, etc.

The process steps of our proposed model "DataSpeak" is given below:

- Retrieve the dataset from the repository using Spark.
- Arrange the whole dataset in increasing order
- Choose the value of k = Number of element in Quartile-1 (i.e., Q1) and store the value in dataset.
- Arrange the calculated sample distance for all quartile training sample. If (xi, yi) is test set data, then Dist(D) =

$$(|a - xi| + |b - yi|)$$

- Sort the sample distances of all quartile and find the nearest neighbors with the set on minimum distance (value K).
- Collect the family of all the nearest neighbors.
- Find predicted score of all the query objects.
- Classify the data on the basis of prediction category.
- All quartile values as Q1, Q2, Q3, and Q4 give the extracted and aggregated data.
- On the basis of aggregated data, we can classify the big data repository, i.e., it is Structured, Unstructured, Semi-Structured, Quasi-Structured, or Quasi-Unstructured.

We are emphasizing on huge training set and evaluating the whole procedure with modified kNN. In case of distance calculated between test data and the training data is also calculated in easy steps. The proposed algorithm of Data Speak is given in the Algorithm 1.

Algorithm 1: DataSpeak: Modified kNN Algorithm with Data Extraction

```
Data: k, Xi, Yi
Result: Classification of Dataset and its Value
initialization: k, Q;
while Calculating the sample query distance do
    Start Spark for retrieval;
    Query distance with Quartile;
    Set data as increased order;
    Data extracted;
while Retrieve the data set using spark do
    if (Q1==k then
        Sample distance of Q1;
        Data retrieval start;
        Category of nearest neighbors;
        Predicted classification of query;
        Data extracted;
        else if (Q2==k) then
            Sample distance of Q2;
            Data retrieval start;
            Category of nearest neighbors;
            Predicted classification of query;
            Data extracted;
            else if (Q3==k) then
                Sample distance of Q3;
                Data retrieval start;
                Category of nearest neighbors;
                Predicted classification of query;
                Data extracted;
    else
        Sample distance of Q4;
        Data retrieval start;
        Category of nearest neighbors;
        Predicted classification of query;
        Data extracted;
```

6 DataSpeak: Dataset Collection

We have executed and analyzed our proposed framework on the huge dataset from travel and tourism [24], placement papers [25], movies and historical [26], smartphone [27], geographical [28], satellite [29], genetic [30], click streamed [27], etc.

Table 1 DataSpeak: dataset for analysis

Dataset sources	Dataset category	Size of data (Mb)
Knoema	Travel and tourism	230
Vyoms	Placement papers	150
IMDB	Movies and historical	360
Google	Smartphone	256
Rtwilson	Geographical	310
Google earth engine	Satellite	243
KDD	Genetic	156
Google	Click streamed	196

We have given fast accessing of data with efficient data extraction in a minimal time when compared to existing algorithm in same domain using same dataset. As concerned with the data aggregation and classification, our approach returns 98% of data aggregation and classification based on data structure using the following dataset collection. See Table 1.

7 DataSpeak: Result and Analysis

As for the result of "DataSpeak", we have used the proposed modified kNN algorithm with Spark framework. Our approach gives fast accessing of data with efficient data extraction and classification in a minimal time when compared to existing kNN and modified algorithms concept. As concerned with the data aggregation and classification, our approach returns fast processing of data aggregation and classification when compared to many proposed researches. See Table 2 and Fig. 4a.

To justify the training and test set, we use to check the precision and recall on the basis of tenfold-based cross validation, the whole dataset is split randomly into ten subsamples. These subsamples (A1–A10) are equally divided. Out of ten sub samples, nine are taken for training set and the rest of subsample is taken as a test set. See Table 3 and Fig. 4b.

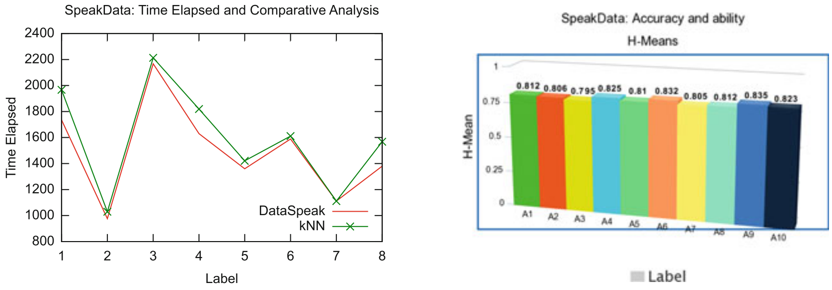
For the classification and aggregation of data structure, we have maintained a confusion matrix. Confusion matrix creates a comparative analysis of data extraction and produces an output in the form of data classification and aggregation. See Table 4.

Table 2 DataSpeak: time elapsed and comparative analysis

Dataset sources	Dataset category	Size of data (Mb)	DataSpeak:Time elapsed (ms)	MkNN or kNN:Time elapsed (ms)	Label
Knoema	Travel and tourism	230	1736	1967	A
Vyoms	Placement papers	150	976	1030	B
IMDB	Movies and historical	360	2167	2213	C
Google	Smartphone	256	1630	1819	D
Rtwilson	Geographical	310	1360	1421	E
Google earth engine	Satellite	243	1589	1611	F
KDD	Genetic	156	1043	1112	G
Google	Click streamed	196	1380	1568	H

Table 3 DataSpeak: accuracy and ability (10-fold)

Group	Test set	Training set	Recall	Precision	H-Mean
A1	A1	A2–A10	0.81	0.82	0.812
A2	A2	A1, A3–A10	0.79	0.80	0.806
A3	A3	A1–A2, A4–A10	0.78	0.79	0.795
A4	A4	A1–A3, A5–A10	0.81	0.83	0.825
A5	A5	A1–A4, A6–A10	0.81	0.82	0.81
A6	A6	A1–A5, A7–A10	0.82	0.84	0.832
A7	A7	A1–A6, A8–A10	0.79	0.81	0.805
A8	A7	A1–A6, A8–A10	0.80	0.81	0.812
A9	A8	A1–A7, A9–A10	0.82	0.84	0.835
A10	A9	A1–A8, A10	0.81	0.83	0.823



(a) DataSpeak: Time Elapsed & Comparative Analysis

(b) DataSpeak: Accuracy and Ability

Fig. 4 DataSpeak: visualization

Table 4 DataSpeak: confusion matrix of data classification and aggregation

Dataset	Structure (%)	Unstructured (%)	Semi-Structured (%)	Quasi-Structured (%)	Quasi-Unstructured (%)
Structure	**97**	01	01	0.5	0.5
Unstructured	0.5	**99**	0.1	0.3	0.1
Semi-Structured	01	0.5	**98**	0.3	0.2
Quasi-Structured	01	0.5	0.2	**98**	0.3
Quasi-Unstructured	01	0.1	0.5	0.4	**98**

8 Conclusion and Future Work

Our approach "DataSpeak" is a more robust approach for extraction, classification, and aggregation in big data. We have tested approx 2 GB of dataset and the result with respect to these dataset is approximately 98%. Our approach overcomes the limitation of kNN, and we have checked the accuracy and ability of the proposed work "DataSpeak" with statistical approaches like cross validation, precision, recall, confusion matrix, and comparative analysis. Our approach gives fast accessing of data with efficient data extraction and classification in a minimal time when compared to existing kNN and modified algorithms concept. As concerned with the data aggregation and classification, our approach returns fast processing of data aggregation and classification when compared to many proposed researches. As for confusion matrix, it creates a comparative analysis of data extraction and produces an output in the form of data classification and aggregation with the result of 98% accuracy. In the future, we will work with some extensive analysis on more network dataset.

References

1. Google Cloud and Big Data. https://cloud.google.com/bigquery/ (2016). Accessed 20 Oct 2016
2. Digital Innovation Mobile Big Data. www.digitalinnovationgazette.com/mobile\big\data/ (2017). Accessed 27 Nov 2017
3. Venturebeat Big Data Analytics. www.venturebeat.com/2015/01/22/big-data-and-mobile-analytics-ready-to-rule-2015/ (2017). Accessed 15 Oct 2017
4. Knowledge Hut Types of Big Data. https://www.knowledgehut.com/blog/bigdata-hadoop/types-of-big-data (2017). Accessed 22 Sept 2017
5. Impact Radius The seven Vs. https://www.impactradius.com/blog/7-vs-big-data/ (2017). Accessed 02 Nov 2017
6. Cover, T., Hart, P.: Nearest neighbor pattern classification. In: IEEE Transactions on Information Theory, vol. 13, Issue 1, pp. 21–27, Jan 1967. https://doi.org/10.1109/TIT.1967.1053964
7. Tang, J., Liu, J., Zhang, M., Mei, Q.: Visualizing large-scale and high-dimensional data. In: Proceedings of the 25th International Conference on WWW, pp. 287–297 (2016)
8. Shankar, V.G., Somani, G., Gaur, M.S., Laxmi, V., Conti, M.: AndroTaint: an efficient android malware detection framework using dynamic taint analysis. In: 2017 ISEA Asia Security and Privacy (ISEASP), Surat, pp. 1–13 (2017). https://doi.org/10.1109/iseasp.2017.7976989
9. Shrivastava, A., Verma, V.K., Shankar, V.G.: XTrap: trapping client and server side XSS vulnerability. In: 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, pp. 394–398 (2016). https://doi.org/10.1109/pdgc.2016.7913227
10. Data Aspirant k- Nearest Neighbor. http://dataaspirant.com/2016/12/23/k-nearest-neighbor-classifier-intro/ (2017). Accessed 14 Aug 2017
11. Yianilos, P.N.: Data structures and algorithms for nearest neighbor search in general metric spaces. In: Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, vol. 93, pp. 311–321 (1993)
12. Vaidya, P.M.: An o(nlogn) algorithm for the all-nearest-neighbors problem. In: Discrete Computational Geometry, vol. 4(2), pp. 101–115 (1989)
13. Apache Apache Spark. https://spark.apache.org/ (2017). Accessed 26 Aug 2017
14. Nada, E., Ahmed, E.: Big data analytics: a literature review paper. In: Lecture Notes in Computer Science, pp. 214–227. Springer, Aug 2014
15. Demetrios, Z.Y., Shonali, K.: Mobile big data analytics: research, practice, and opportunities. In: Proceeding MDM' 2014, 15th International Conference on Mobile Data Management, vol. 01, pp. 1–2 (2014)
16. He, Y., Yu, F.R., Zhao, N., Yin, H., Yao, H., Robert, C.: Big data analytics in mobile cellular networks. In: IEEE Access, vol. 4 (2016). https://doi.org/10.1109/access.2016.2540520
17. EMC.: Dell EMC data science analytics. In: EMC Education Services, pp. 1–508 (2015)
18. Shankar, V.G., Somani, G.: Anti-Hijack: runtime detection of malware initiated hijacking in android. In: Procedia Computer Science, vol. 78, pp. 587–594 (2016). https://doi.org/10.1016/j.procs.2016.02.105
19. Fu, C., Cai, D.: EFANNA: An extremely fast approximate nearest neighbor search algorithm based on kNN graph. In: Computer Vision and Pattern Recognition (2016). http://arxiv.org/abs/1609.07228
20. Georgios, S., Mavromoustakis, C.X., Mastorakis, G., Batalla, J.M., Dobre, C., Panagiotakis, S., Pallis, E.: Big data and cloud computing: a survey of the state-of-the-art and research challenges. In: Advances in Mobile Cloud Computing and Big Data in the 5G Era Studies in Big Data 22 (2017)
21. Kune, R., Konugurthi, P.K., Agarwal, A., Chillarige, R.R., Buyya, R.: The anatomy of big data computing. In: Softw. Pract. Exper. **46**, 79105 (2016)
22. Yang, C., Huang, Q., Li, Z., Liu, K., Hu, F.: Big Data and cloud computing: innovation opportunities and challenges. In: International Journal of Digital Earth. Published by Informa UK Limited, trading as Taylor Francis (2016)
23. Tsai, C.W., Lai, C.F., Chao, H.C., Vasilakos, A.V.: Big Data Anal Surv J. Big Data **2**, 21 (2015). https://doi.org/10.1186/s40537-015-0030-3

24. Knoema Tourism Dataset. https://knoema.com/atlas/topics/Tourism/datasets (2017). Accessed 24 Oct 2017
25. Vyoms Placement Dataset. http://www.vyoms.com/placement-papers/domains/details/business-analysis-223.asp (2017). Accessed 28 Oct 2017
26. IMDB Movies Dataset. https://www.kaggle.com/orgesleka/imdbmovies (2017). Accessed 21 Oct 2017
27. Google Smartphone Dataset. https://cloud.google.com/public-datasets/ (2017). Accessed 19 Oct 2017
28. Rtwilson Geographical Dataset. https://freegisdata.rtwilson.com/ (2017). Accessed 13 Oct 2017
29. Google Satellite Dataset. https://earthengine.google.com/datasets/ (2017). Accessed 13 Oct 2017
30. MIT Genetic Dataset. https://www.ll.mit.edu//ideval/data/ (2017). Accessed 17 Oct 2017
31. Shankar, V.G., Jangid, M., Devi, B., Kabra, S.: Mobile big data: malware and its analysis. In: Proceedings of First International Conference on Smart System, Innovations and Computing. Smart Innovation, Systems and Technologies, vol. 79, pp. 831–842, Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-5828-8_79

Design and Implementation of Internet of Things Based Multi-sensor Device

Ravikant Khamitkar and Farid Valsangkar

Abstract As the technology use in the world progresses the multiple devices and things will co-operate with each other to achieve high reliability and accuracy in the information sharing and processing for better future. To achieve this, the integration of various sensors in single device is designed with the help of standard enclosure and wireless communication technology. The device consists of various physical parameter sensors like temperature, humidity, light intensity, proximity are integrated in single enclosure which communicates its parameters over wireless standard 802.11 a/b/g/n. The access to the information and or sensor data is achieved using Secure Message Queuing Telemetry Transport (SMQTT) standard protocol. The Multi-sensor Device have 32-bit controller based chip ESP8266 programmed to establish the communication with other devices using IBMs Node-Red programming tool running on Raspberry Pi. The data is logged locally on the memory of Raspberry pi. The data is accessible from outside network with secured authentication. The display of sensor data is done on the Node Red based user interface for easy access.

Keywords SMQTT · Node-Red · Raspberry pi

1 Introduction

With technological advancement for information acquisition and analysis it is necessary to use it for optimum benefits of mankind. The things and devices getting connected to each other through the wireless and wired channel, the networking of these devices through the Internet gives access and control effortlessly. The use of multiple sensors for information acquisition is required many times for various appli-

R. Khamitkar (✉) · F. Valsangkar
N.B.N. Sinhgad College of Engineering, Solapur 413003, MH, India
e-mail: khamitkar.ravikant@gmail.com

F. Valsangkar
e-mail: valsangkarfarid410@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_17

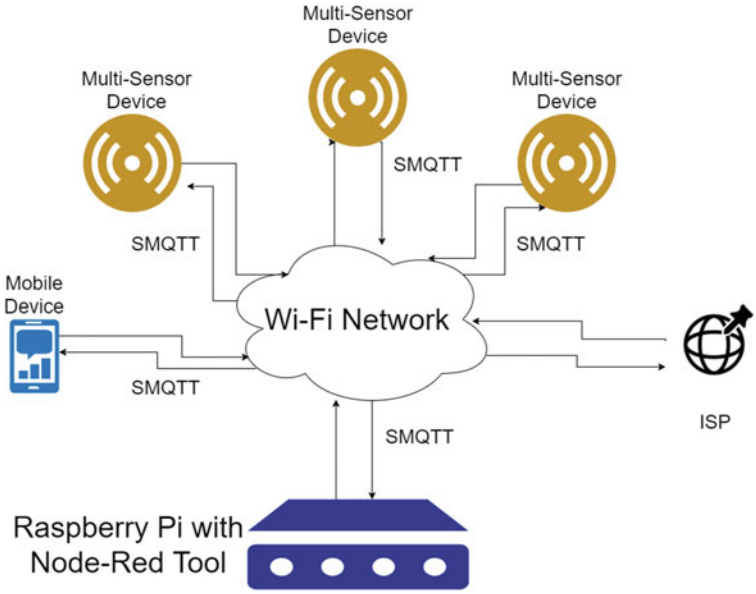


Fig. 1 Multi-sensor system overview

cations. The integration of multiple sensors in single package with communication through the wireless and wired channel can be achieved using various protocols like SMQTT, AMQP protocol. The SMQTT is used for the implementation of practical setup of multi-sensor device [1, 2]. The coding is done using Node-Red programming tool running on Raspberry Pi platform. The Sensors are connected to ESP8266 having 32 bit controller and it communicates with Raspberry Pi over wireless channel using IEEE 802.11 standard. The user interface is designed using Node-Red user interface and is accessed using IP address of Raspberry pi Device running the SMQTT broker and Node Red tool.

2 Hardware Overview

The overall system is having the two parts one is Multi-Sensor nodes (ESP8266) and central node (Raspberry Pi) with Node-Red tool. As shown in figure it can be seen that the multi-sensor devices are connected to raspberry pi based SMQTT broker (Mosquitto) using Wi-Fi network [3]. The ISP is also connected to Wi-Fi network for internet connectivity. The mobile device is also connected to Wi-Fi network and it can access the various devices sensor data using Node-Red based user interface (Fig. 1).

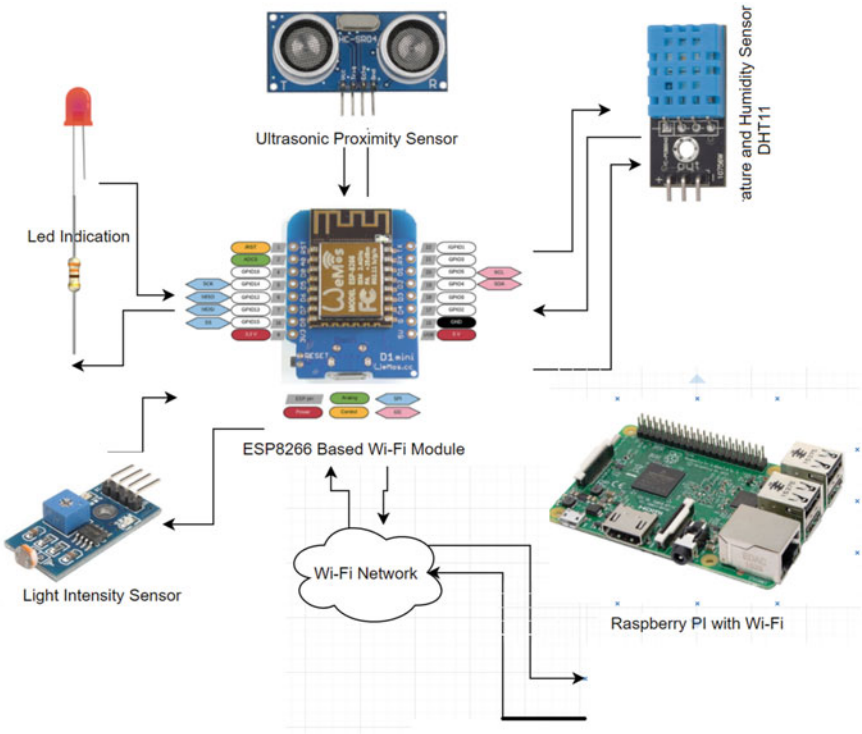


Fig. 2 Architecture of Multi-sensor device

The individual multi-sensor device consists of ESP8266 connected with one pin Temperature and humidity sensor (DHT11), Ultrasonic Sensor for proximity detection, Light Intensity Sensor (LDR) and indication LED. The DHT11 is connected to ESP8266 board through single pin 2 and *vcc* and *gnd* pins are connected on the board. The Ultrasonic sensor is connected to pins 13 (Trigger) and 12 (Echo) on board. The Light intensity sensor is connected to pin A0 of the ESP8266 board. The LED is interfaced to board via pin 5 for indication purpose. The Board communicates with Raspberry pi based SMQTT broker and Node-Red tool over Wi-Fi network. The Raspbian OS is running on Raspberry pi with mosquitto (MQTT) broker and the IBM Node-Red programming tool [4]. This acts as central node for communicating with different multi-sensor devices. The Raspberry pi locally stores the data from various sensors of different nodes and is also visible and available for different users via the Node-Red based user interface (Fig. 2).

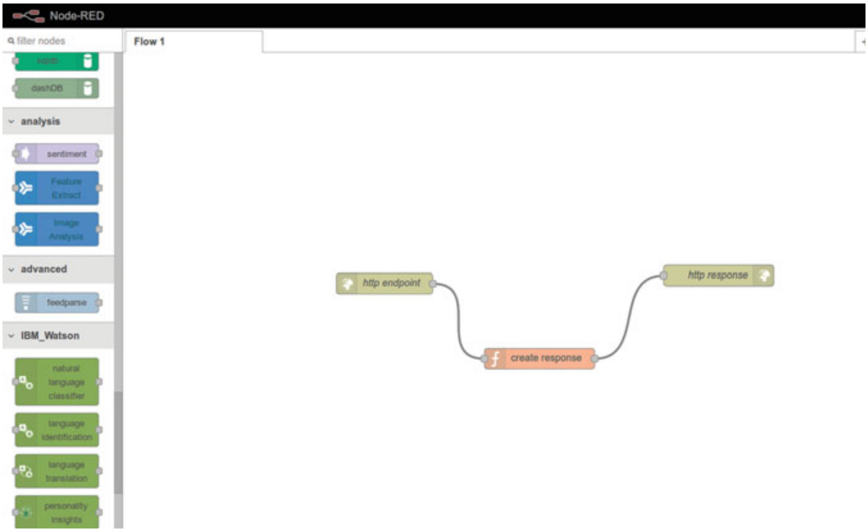


Fig. 3 Node Red based system design

Table 1 SMQTT topics, publishers and subscribers

Topic name	Publisher	Subscriber
Room/lamp	Switch	LED on board
Room/temperature	DHT11 (Temp.)	UI Chart
Room/humidity	DHT11 (Humidity)	UI Gauge
Room/light	Light Intensity Sensor	UI Gauge
Room/proxydist	Ultrasonic Sensor (Distance)	UI Bar Graph

3 Software Overview

The software used for the design of the system consists of Node-Red programming tool, Mosquitto SMQTT broker running on Raspberry Pi and the Arduino IDE for ESP8266 board programming. The ESP8266 is programmed to collect the data from the various sensors and publish the data through SMQTT broker via particular topics. The Raspberry pi based Node-red have user interface which subscribe to the topics of SMQTT broker to collect and display the sensor data on screen.

From the Fig. 3 it can be seen that there are five MQTT topic nodes which are connected to the user interface nodes. The data coming from four different sensors is combined and given to the sensor database file. This file stores the data locally [5–7]. The database also can be used to access through the cloud like blinky app, maker.io, etc.

The SMQTT protocol have total five topics as listed below (Table 1).

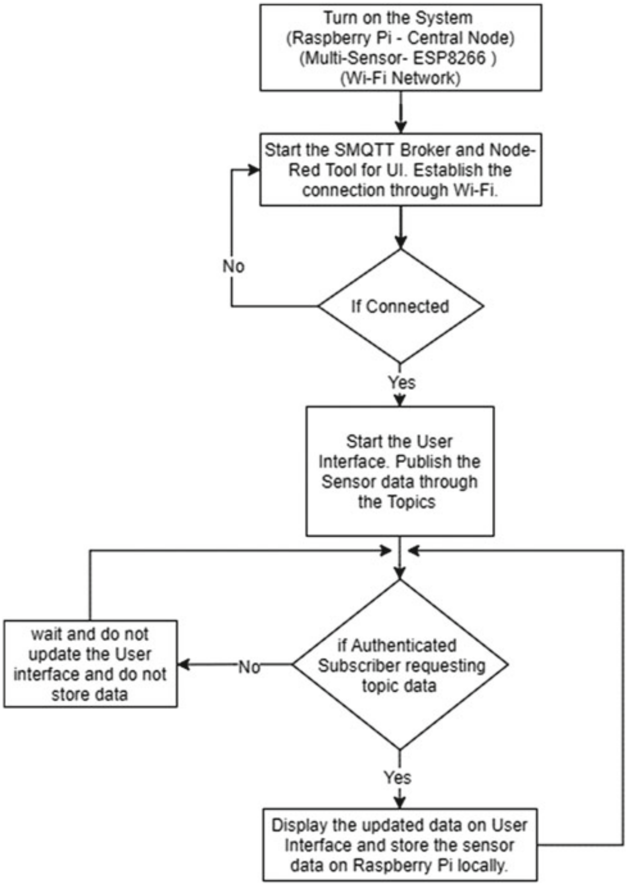


Fig. 4 Flow of software interface

The Security is used for the Wi-Fi network access via the SSID and Password encryption. The User interface and administrator access for Node-Red is allowed through authentication process only. This helps to avoid unauthorized access to the sensor data and its use.

Flowchart for the system work is as follows (Fig. 4).

4 Result and Discussion

The system gives the sensor data output display on the user interface for the authenticated users only. The sensors connected to ESP8266 board sends the data to the

Table 2 Sensor parameters range

Sensor name	Min	Max	Unit
Temperature	0	80	°C
Humidity	0	100	%
Ultrasonic sensor (Proximity)	0	300	CM
Light intensity	0	100	%

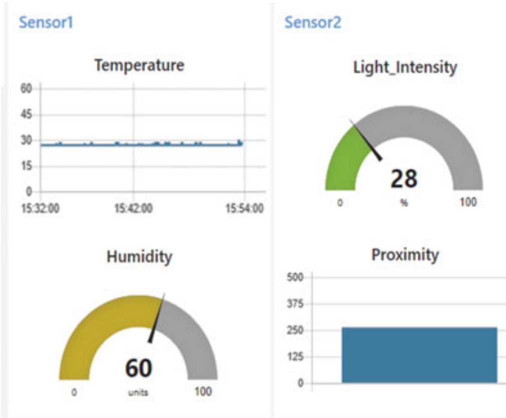


Fig. 5 User interface for the sensor device

Table 3 Comparison with Arduino Based System

Particular	Arduino based system	ESP based multi sensor device
Base device used	Arduino	ESP8266
Working frequency (MHz)	16	80
Number of sensors connected	4	4
Protocol used for communication of DATA	GSM/GPRS	Wi-Fi/SMQTT
Model of communication	Client-Server	Publish-Subscribe

central node consists of Raspberry pi [4]. The low to high range of the different Sensor Parameters is as shown in the Table 2.

The User interface is shown below consists of a Switch for led (lamp) control. There are two groups of Sensors for display as Sensor1 and Sensor2. Sensor1 have Temperature and Humidity display. And Sensor2 have Light intensity and Proximity display. The user interface is updated every three seconds (Fig. 5).

The data is stored on the raspberry pi using Local file storage node in the configuration of the system. The comparison of Multi sensor device is done with the system based on Arduino in [4] as based on the following parameters (Table 3).

5 Conclusion

From the implementation of multi-sensor device it can be concluded that the Node Red platform provides customizable programming options for configuration. The options for selecting the communication protocol like SMQTT, AMQP, etc., gives flexibility in design. The security can be added for authentication as well as message passing over the end-to-end network. The easy replication of the multiple device makes it scalable. The cost-effective implementation is possible for the common use. The implementation cost of multi-sensor device will be less as compared with the device off-the-shelf available in market with same number of sensors.

References

1. Jhon, R., Cesar, P.: Performance analysis of communication protocols for Internet of Things Platforms. In: 2017 IEEE Conference Communications and Computing (COLCOM) (2017). https://doi.org/10.1109/colcomcon.2017.8088198
2. Naik, N.: Choice of effective messaging protocols for IoT systems: MQTT, CoAP AMQP and HTTP. Syst. Eng. Symp. (ISSE) (2017). https://doi.org/10.1109/SysEng.2017.8088251
3. Prathibha, S.R., Anupama, H., Jyothi, M.P.: IOT Based monitoring system in smart agriculture. Recent Adv. Electron. Commun. Technol. (ICRAECT) (2017). https://doi.org/10.1109/icraect.2017.52
4. Hazem, M.R., Hossain, M.S.: Fog intelligence for real-time IoT sensor data analytics. In: IEEE ACCESS, vol. XX, no. X, Sept 2018. https://doi.org/10.1109/access.2017.2754538
5. Angelo, C., Francesco, P., Giuseppe, R.: SMuNe a smart multisensor network based on embedded systems in IoT environment. Signal-Image Technol. Internet-Based Syst. (SITIS) (2015). https://doi.org/10.1109/SITIS.2015.51
6. Zenon, C., Robin, B.: Learning data engineering: Creating IOT apps using the node-RED and the RPI technologies. Inf. Technol. Based High. Educ. Train. (ITHET) (2017). https://doi.org/10.1109/ITHET.2017.8067827
7. Running Node-Red on Raspberry Pi (2017). https://nodered.org/docs/hardware/raspberrypi

Internet of Things for Irrigation Monitoring and Controlling

R. J. Muley and V. N. Bhonge

Abstract The three things that the world runs on are water, wheel, and fire. Agriculture is largely based on water, its proper usage, storage, and management. This is a very key issue. With the advent of technology, manual systems evolved from being mechanized to being automatic and recently to being smart. The technology wave of Internet of Things can further ease the tedious task of watering and irrigating the fields, especially in the water scarce regions as well as in situations where fields are far away from residential areas. Also it saves the excess use of water for any crop. Erratic climate behavior in the Indian subcontinent in the last few years has led to extreme water scarcity in many regions. This paper proposes an irrigation monitoring and controlling system based on Wireless Sensor Network (WSN) and Internet of Things. The WSN remotely collects the data from the fields and transfers it to the cloud. The wireless sensor network uses two sensors: DHT and soil moisture sensor. A management server accesses the information over the cloud, a graphics user interface processes it and generates a feedback. This feedback is based on a user-selected crop name (stored in database) in the GUI. The information is sent over the cloud via IoT gateway. In this work, the hotspot from the mobile phone is used as an IoT gateway. The ESP8266-12E is used as controller as well as a Wi-Fi module. The feedback device is a 12 V DC pump. Whenever the soil moisture level senses dry soil (0-min %), the management system will generate a feedback signal to switch on the pump.

Keywords Internet of things (IoT) · Wireless sensor network (WSN) · IoT gateway

R. J. Muley (✉) · V. N. Bhonge
Department of Electronics and Telecommunication, SSGMCE, Shegaon, India
e-mail: rjm1430007@gmail.com

V. N. Bhonge
e-mail: vnbhonge@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_18

1 Introduction

Agriculture is the livelihood of people in many countries. In recent years food scarcity is increasing so productivity and food security are important issues. To increase productivity more precise farming technique and support from environmental conditions plays crucial role. Farmers face problems such as fertilizer abuse, diseases on crop, reduction in arable land, water waste, fragmentation of land, etc. [1]. Fertilizer abuse and diseases on crop can be stopped by providing ample of fertilizers on loan basis at the time of plantation and providing insecticides and germicides at concessional rates. Fragmentation of land should stop if issues on farm land are solved by mutual understandings. By utilizing arable land periodically, its infertility can be avoided. Water waste needs special attention, since rain water needs to be stored other than natural resources and wastage of water also occur while watering the farm, so usage of water should be done properly. Due to mismanagement of water resources many areas are facing drought like situations. Water resources are not able to provide sufficient water to farms over duration of a year. In traditional technique, farmers supply water periodically, he remains present during watering hours even at night. There may be situations of irregular supply, more or less supply of water, which can affect the development of crop. Irrigation can be regulated by automation in farming technique. An automated system can be installed in farms which will continuously monitor the moisture level of soil and according to need water will be supplied. One of the applications in irrigation was started with SMS commanding system where based on soil humidity water was supplied [2], with a disadvantage that farmers were not informed about the irrigation status and it did not consider the condition or type of soil.

Another solution was proposed based on Wireless Sensor Network (WSN) [3]. WSN involves sensors, communication modules and data processing modules, which monitors soil parameters remotely. It has disadvantage that monitoring can be realized at short distance by constructing wireless network in ad hoc manner. Data from WSN cannot be sent over long distance, because it lacks uniform standard communication protocol.

The third solution is the newer one based on WSN and has capability to transmit data using Internet anywhere, an Internet of Things (IoT). IoT bridges the gap between communication technologies and sensor network so that network communication can be done easier and network devices can be managed. IoT has proved a real time solution to many applications such as home automation, healthcare, industrial Internet, weather [4] etc. With the advent of IoT, it has become very easy to handle regular task on just a click. It makes use of growing internet technology more efficiently and created work environment where internet devices are connected to smart phones [5]. As IoT takes help of WSN, sensors installed on farm can send data to farmer even if he is not able to reach the farm. With this data available, he can send command to system which will operate accordingly [6].

Many researchers are working in this field, Rajlakshmi and Mahalakshmi [7] proposed IoT-based crop field monitoring and smart irrigation automation system which communicate wirelessly from field to famers and conveys real-time information of farm. System includes soil moisture, temperature, humidity and light sensors which capture data and processed by Arduino microcontroller. The data when reaches to farmer, decision is made to turn ON or OFF the irrigation system. Light sensor, LDR is used to know the lighting status so that photosynthesis action input can be known. NRF24L01 wireless module is used to transfer data over wireless by Zigbee and GPS/GPRS technology. Farmer is provided with web-based application to monitor the farm data. Bhanu et al. [8] proposed WSN based monitoring of soil parameters for effective irrigation, according to soil type water holding capacity varies so authors have provided solution to monitor soil water content. Humidity sensors senses the water content and data is sent through wireless network to server and then to farmer so that irrigation event can be started or stopped. WSN system mainly consists of sensors, transreceivers and microcontroller devices. Abedin et al. [9] worked on IP based WSN for smart irrigation systems; it uses architecture of WSN and IoT, which analyses moisture level in the field. Communication from farm is done by 6LoWPAN, RPL and IEEE 802.15.4 g protocols. Cambra et al. [6] proposed an IoT Service Oriented Architectures (SoA) for agriculture monitoring; it uses aerial images along with IoT and SOA so that it can be applied to large farms. The system presents multimedia platform for precise agriculture. The farm region is monitored by AR Drones and irrigation scheduling is done via IoT. Such system covers an area of 12 km. Khelifa et al. [10] proposed smart irrigation system in Algerian environment. Irrigation system is operated from long distance using 6LoWPAN WSN which connects the Zigbee network with Internet. Pump is operated to turn ON or OFF the water supply to farm field.

2 Architecture of Internet of Things for Irrigation Monitoring and Controlling

The architecture of proposed system has three parts as shown in Fig. 1:

1. Wireless Sensor Node
2. IoT Gateway
3. Management System: PC with Internet (software GUI).

The functions of each part is as explained below:

1. **Wireless Sensor Node (WSN)**:

There are four WSN in the system. Each WSN comprises of two sensors as input and a 12 V DC pump as output. The input sensors are: DHT11, soil moisture sensor. These things are connected to the internet by a Wi-Fi module-controller: the ESP8266 12-E. Each is explained in detail as follows:

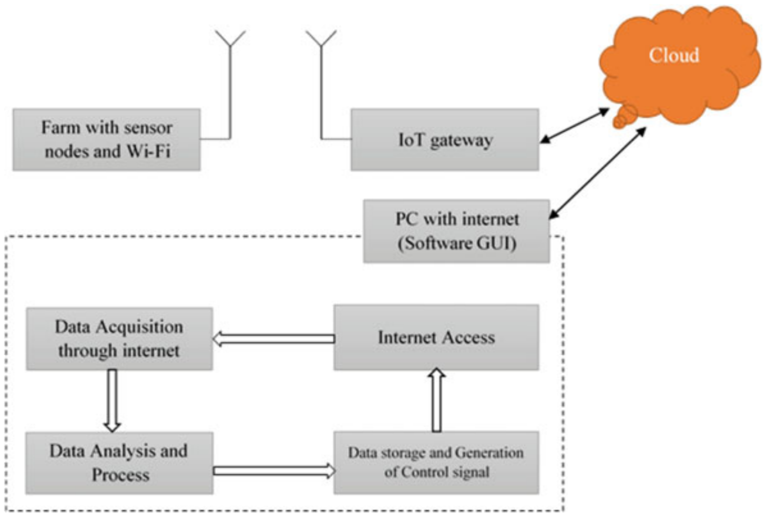


Fig. 1 Architecture of proposed system

A. **Soil Moisture Sensor**:

This sensor collects the moisture data from soil. If the soil is dry, it will be much resistive and hence the value shown will be high. The comparator shows these values from 0 to 1023 levels. The more the value tends towards 1023, the dry will be the soil [11].

B. **DHT–11**:

DHT-11 is useful to measure the atmospheric temperature and humidity. This sensor calculates the relative humidity of the air [15]. In this work, temperature near the roots is considered. The sensors are mounted on the acrylic sheet and at a height of 1 inch above the ground level.

C. **Wi-Fi Networking Solution**:

ESP8266-12E [12] offers a complete and self-contained Wi-Fi networking solution. For this system, it is configured as a client. It is programmed using Arduino IDE.

D. **Pump**:

A feedback is generated to the output device: a 12 V DC submersible pump [13], used for model purpose. For commercial applications, it can be changed as per requirement of user.

These three things (DHT, soil moisture sensor, and pump) upload the data on the Dweet URL created for each thing.

2. **IoT Gateway**:

The path over which the data is sent or received over the cloud is called as the IoT gateway. This could be any general router. The hotspot from mobile phone is used as

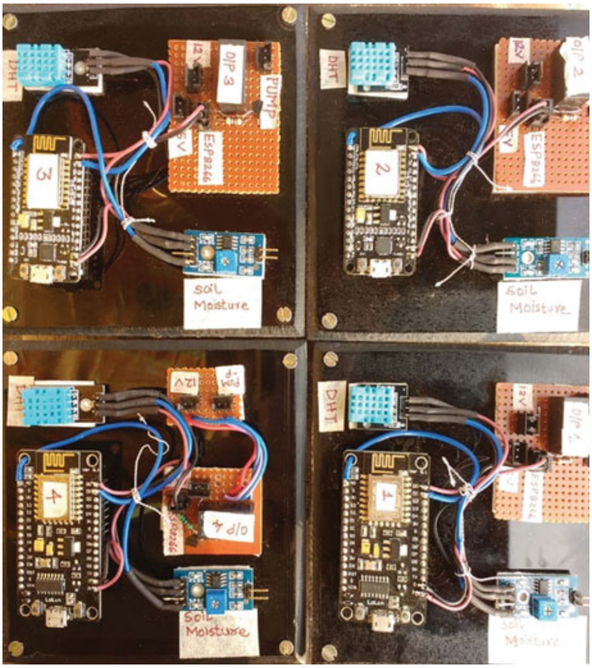


Fig. 2 Hardware implementation of WSN

an IoT gateway. The model of the mobile is MOTO G (Third Generation). It provides the hotspot with an SSID and password at a 2.4 GHz band.

3. **Management Server**:

The management GUI-A Graphics User Interface (GUI) is designed using open source application Software PROCESSING. It is an open source project initiated by Casey Reas [14]. Presently it operates under the non-profit Processing Foundation, 501(c) (3). The PROCESSING Software is a flexible software sketchbook and a language for learning how to code within the context of visual arts. The PROCESSING software is free and open source and runs on MAC, WINDOWS and GNU/LINUX platforms. It is used to monitor and control the WSN.

Figure 2 shows the implemented hardware. Each WSN is numbered 1, 2, 3 and 4. Each is implanted on four parts of the field.

3 Methodology

Figure 3 shows flowchart of the working of system for one cycle. Switching ON the system, the ESP8266-12E collects data from two sensors. The data collected at an interval of 1 s, is temperature, humidity and soil moisture. ESP8266-12E collects data from the sensors and sends to their destined Dweet API via IoT gateway. Dweet

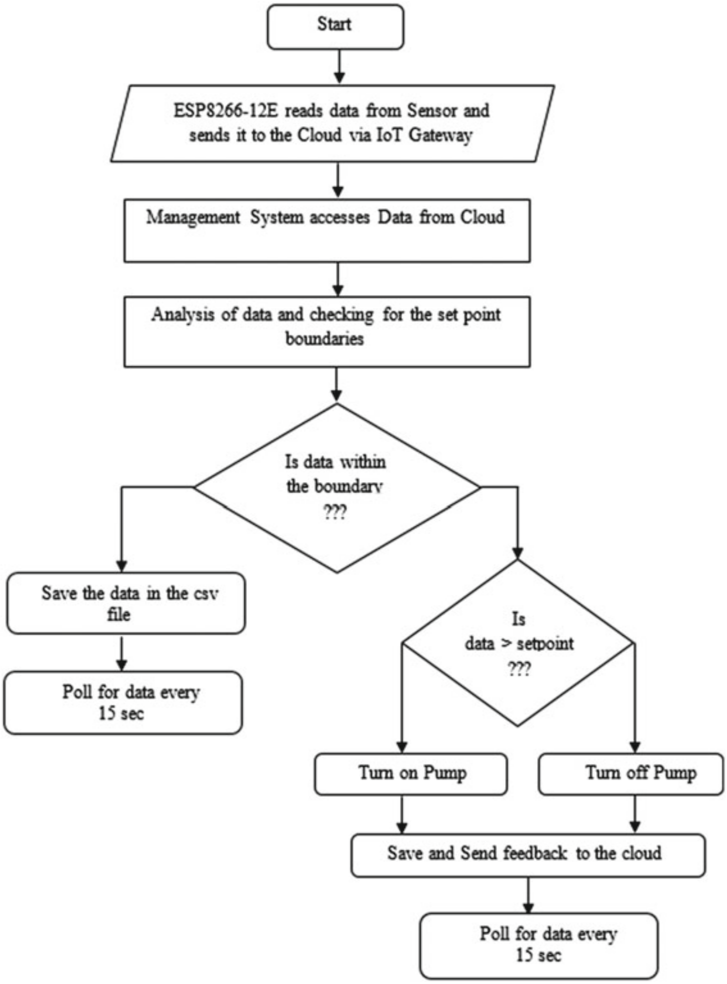


Fig. 3 Flowchart of methodology

is a messaging and alert system for IoT. It is fast, free and easy to use. ‘Dweeting’ means sending data from the thing to the cloud with web API. User can configure the Dweet API console. The information on Dweet is safe and secure, due to its locking provision. It is accessed by the management system, a decision is taken based on the database minimum and maximum. The decision is either to switch ON or switch OFF the pump. It is then again uploaded to the destined Dweet API. This upload is accessed by the destined WSN and action is initiated. Following are the URL that can be used to retrieve the Dweets for this system:

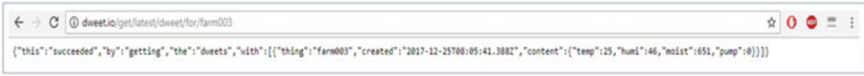


Fig. 4 Dweet result for node 3

For node 1: http://dweet.io/get/latest/dweet/for/farm001
For node 2: http://dweet.io/get/latest/dweet/for/farm002
For node 3: http://dweet.io/get/latest/dweet/for/farm003
For node 4: http://dweet.io/get/latest/dweet/for/farm004

Figure 4 shows the results from above Dweet URL.

For pump 1: http://dweet.io/get/latest/dweet/for/pump001
For pump 2: http://dweet.io/get/latest/dweet/for/pump002
For pump 3: http://dweet.io/get/latest/dweet/for/pump003
For pump 4: http://dweet.io/get/latest/dweet/for/pump004

These APIs will be accessed by the respective WSN as well as the management system. To access or view these Dweets, simply enter the URL in browser and the Dweets for last 24 h will be shown. If there are no Dweets in the last 24 h, then the browser shows the following message:

{"this":"failed","with":404,"because":"we couldn't find this"}

4 Results

To monitor and control the irrigation activity, a database is created according to crop and its minimum and maximum soil moisture level. The database is as mentioned in Table 1. The Dweet results are explained as follows: from Fig. 4, the soil moisture reading of node 3 is very high; i.e., more than maximum limit for Jowar crop. It shows that the soil is moist enough as required for the crop. Hence, the pump is turned off. Similarly that will be the case of nodes 3 and 4. From Fig. 5, the soil moisture level of node 1 is below minimum for the crop hence the pump is turned ON. The graphs in the right side show the temperature, humidity and moisture readings for all the four nodes with respective colors as specified in the bottom.

Table 1 Database

Crop	Min	Max
Rice	22	32
Wheat	25	26
Cotton	20	50
Bajra	21	63
Barely	18	58
Jowar	18	55
Beans	23	32
Citrus	23	32
Maize	25	51
Sorghum	23	65
Soyabean	14	45
Sunflower	32	50
Apricots	52	80

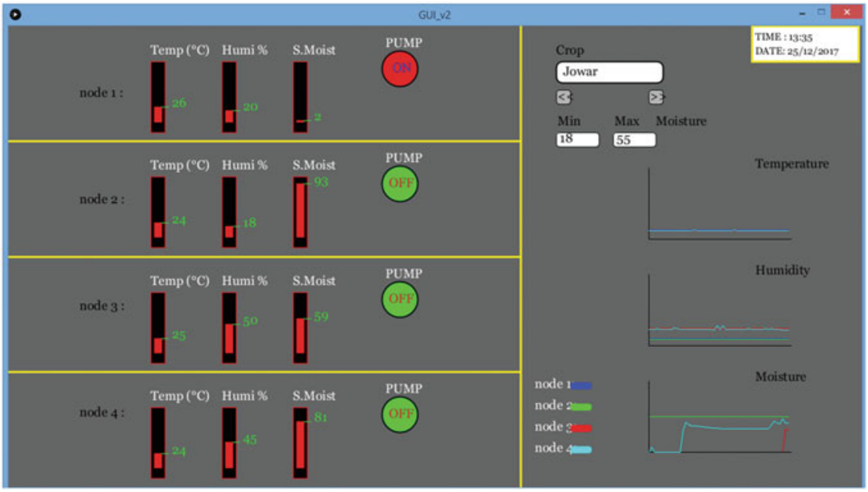


Fig. 5 GUI developed to monitor and control Irrigation activity

5 Conclusion

In this work, irrigation monitoring and controlling is implemented with WSN and IoT technology. Also, an effective solution is provided to monitor the three basic parameters of field: temperature, humidity, and soil moisture. The management system with feedback is designed appropriately. The whole system has the advantage of small volume and high reliability. This system brings innovation to the existing technology in the irrigation and improves the safety features, hence proving to be an effective development in the agriculture industry. Irrigation monitoring system confirms that the field is not over irrigated and all the data is well recorded for prediction purpose. In future, sensors and actuators can be added as per the need of user and application.

References

1. Sawant, S., Adinarayana, J., Durbha, S.: Krishisense: a semantically aware web enabled wireless sensor network system for precision agriculture applications. In: IEEE Geoscience and Remote Sensing Symposium, pp. 4090–4093(2014)
2. Patil, K., Kale, R.: A model for smart agriculture using IoT. In: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGT-SPICC), pp. 543–545 (2016)
3. Harun, A., Kassim, M., Mat, I., Ramli, S.: Precision irrigation using Wireless Sensor Network. In: 2015 International Conference on Smart Sensors and Applications (ICSSA), pp. 71–75(2015)
4. Deshpande, P., Iyer, B.: Research directions in the internet of every things (IoET). In: International Conference on Computing, Communication and Automation (ICCCA), pp. 1353–1357 (2017)
5. Parmeswaran, G., Sivaprasath, K.: Arduino based smart drip irrigation system using internet of things. Int. J. Eng. Sci. Comput. **6**(5), 5518–5521 (2016)
6. Cambra, C., Sendra, S., Lloret, J., Garcia, L.: An IoT service-oriented system for agriculture monitoring. In: IEEE International Conference on Communications, pp. 1–6 (2017)
7. Rajalakshmi, P., Mahalakshmi, S.: IOT based crop-field monitoring and irrigation automation. In: 10th International Conference on Intelligent Systems and Control (ISCO), pp. 1–6 (2016)
8. Bhanu, B., Hussain, H., Ande, P.: Monitoring of Soil Parameters for Effective Irrigation using WSN, International Conference on Advanced Computing (ICoAC), 211– 2015 (2014)
9. Abedin, M., Chowdhury, A., Hossain, M., Anderson, K., Karaim, R.: An interoperable IP based WSN for smart irrigation systems. In: IEEE Annual Consumer Communications and Networking Conference (CCNC), pp. 1–5 (2017)
10. Khelifa, B., Amel, D., Amel, B., Chabane, M., Tarek, B.: In: Smart Irrigation using Internet of Things, International Conference on Future Generation Communication Technologies (FGCT), pp. 91–96 (2015)
11. A soil moisture sensor Datasheet, http://forum.researchdesignlab.com/datasheet/sensors/soil%20moisture%20sensor.pdf
12. ESP8266-12E Datasheet. http://www.kloppenborg.net/images/blog/esp8266/esp8266-esp12e-specs.pdf

13. A 12 V DC Submersible Pump Datasheet. http://marchmay.co.uk/includes/pumps/datasheets/0544_893-datasheet.pdf
14. Processing Application Software Version 3. https://processing.org/
15. DHT11 Datasheet. https://akizukidenshi.com/download/ds/aosong/DHT11.pdf

Hostel Rooms Power Management and Monitoring Using Internet of Things

Meenakshi Patil, Vijay D. Chaudhari, Hemraj V. Dhande and H. T. Ingale

Abstract Power saving is the important issue nowadays, and it is more critical in hostels because of some irresponsible students who leave the room without switching OFF the tubes lights and fans. So, for controlling this wastage of electricity in hostels, we have tried here to develop this system that helps in monitoring and managing the electrical power requirement. In this system, IR sensors sensed the presence of students in the room with the help of counter. When it counts one, it means students enters the room, this counter increases and so on. When the student leaves the room, it decreases the count and when it reaches up to zero, this indicates that no one is present in the room. At this time, after few seconds, the switches will automatically turn OFF, if it is ON, and this information will be sent to the server/cloud where the authorized person can see or watch all the activities in the room. Here, we need only Internet for watching the online process. This will be in the form of notification where it will show the room number, OFF time. The other feature of this system is, when such notifications will be seen on the screen, one SMS will be sent to the student about Rs. 100/- penalty or punishment. Internet of Things plays a vital role in this system. This promotes students to become responsible about careful utilization of electricity.

Keywords ATmega 328p · ESP8266EX · Optocoupler · IR sensor · Electrical power · IoT

M. Patil
VLSI & ESD, E & TC Department, GF's Godavari CoE, NMU, Jalgaon, India
e-mail: meenakship1407@gmail.com

V. D. Chaudhari (✉) · H. V. Dhande · H. T. Ingale
E & TC Engineering Department, GF's Godavari CoE, NMU, Jalgaon, India
e-mail: vinuda_chaudhari@yahoo.co.in

H. V. Dhande
e-mail: hemrajd99@gmail.com

H. T. Ingale
e-mail: hetui@rediffmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_19

1 Introduction

Power saving is an increasingly important issue for all hostels providing accommodation for the students. There is number of existing systems, which are used for power saving in the hostels but having limitations like Monitoring System. As most off students tend to forget to turn off the lights or fans on their way out of the room, the Monitoring System controls the wastage of power by monitoring the occupancy in the room and turning ON the lights/fans in every room that is unoccupied. To maintain daily monitoring, the record is complicated and time-consuming for the hostel management. Hostel Rooms Power Management and Monitoring System are proposed to reduce the energy wastage due to some irresponsible students in the hostel. The system makes automatically monitor the status of rooms that is it empty without switched off the tubes lights and fans in the room or how many students present in the hostel room by means of IR sensors. To build this system, we have used the latest software, Arduino which is used for programming and compiling, OrCAD for circuit design. Also, we have used PHP and HTML for web application and for data management. To sense the number of occupancies in the room, we have used IR transmitter and receiver.

2 Related Work

Different authors have proposed their concern method in the field of electrical energy conservation.

The main goals of energy management system are energy cost reduction and energy conservation [1]. This management task is supported by hardware and software information systems. It means that we should control switches of any room from remote control devices [2]. It provides convenience to user in a busy schedule. This system is convenient and helpful to elderly or disabled persons. In 2002, a new distribution automation technology is used to control distribution voltages for reducing energy uses and control demand and regulation [3].

Fuzzy technique is used [4] to control the street lights and lighting the enterprises. Using this technique, the control system can sense the natural light and switches the lights automatically. Earlier, large number of sensors was required for home automation and energy monitoring. Here, only a single sensor has been used [5]. ZigBee controller has been used to control the dimming light effect on our home appliances [6]. Optical multiplexing is used to switch the devices digitally with high speed [7]. To use the energy efficiently, the dynamic switching is used in real time [8]. The energy saving depends on whether the room occupancy or unoccupancy is implemented using multi agent system like wireless or Ethernet network [9]. Han et al. [10] have explained how energy conservation is realized and monitored using ZigBee hub. Microcontroller has been used for efficient energy management in the room using IR remote control [11].

In [12], the author Luigi established an integrated lighting control concept for utilization of improved lighting design practices, improvements for unoccupied and daylight hours. Wei et al. [13], proposed a system to achieve real-time monitoring and control, and improve the energy saving of building intelligently. The wireless sensor and actuator networks (WSANs) has been used for energy management systems [14]. Akpama et al. [15], mentioned that reduction of demand of power through improved efficiency of devices and procedures. The awareness of new technologies and the cost implication of their application of energy efficiency are explained. The gateway has shown how much energy is required and consumed in summer. Bjelica et al. [16], proposed web applications to interact with remote devices. Asynchronous API calls are used. Liang Hsu et al. [17], mentioned continuous consumption of electricity in greenhouse. Control solutions have been introduced to use the indoor energy efficiently [18]. The Energy Management in South Africa is important issue because of budget cuts and increased financial restrictions. This article focuses on the detailed site audit of all the electrical loads at a tertiary institution. The recommendations include various techniques that can be used to reduce, even further, the cost of electrical energy on campus [19]. The case study of university building has been taken for explanation and presents a technique to manage a building power for energy saving. This case study explained schedule events in general purpose classrooms with minimum energy consumption [20]. Baraka et al. [21], used an Arduino microcontroller has been used for home automation. The energy-saving devices can be used when devices are not switched off by the users while leaving the office. The future Internet smart grid application is used to automatically adapt light, heating, ventilation and air conditioning, and power save modes of PCs and laptops to the presence or absence of users [22]. Goyal et al. have explained an energy audit, survey, and analysis of energy flow for energy conservation in an institute [23]. Energy audit at a Hostel has been held to estimate the energy consumption daily, monthly, and annually [24]. Arduino microcontroller based Small Smart Home System using WLAN network is developed for household appliance control [25]. The Arduino-based system records data when there are power failures [26]. The system is responsible for providing real-time clock and date needed in the data logging operation. The system shows display in real time, the consumed average power in watt along with its price [27]. Controlling of home appliances through a single device is explained in this paper [28]. S. Bhuvaneswari et al. proposed a priority concept to switch the electrical appliances, so as to manage electricity effectively [29]. G. Mendez et al., have explained the support system for energy conservation using sensors, actuators, and agents [30].

Dey et al. [31], carried out the study PV-based microgrid to secure the load demands of a university building. The system is modeled in MATLAB Simulink software (SimPower). Reena et al. considered Building Automation System (BAS) for research [32]. Chaphekar et al. explained the technique for reduction in energy cost by reduction of energy consumption. The author has carried out energy audit at College of Engineering, Pune5 (COEP) hostel campus for this study [33]. On the basis of the comprehensive survey given in the paper [34], the authors have tried to implement this system. Recently, IoT-based technology is getting popularity due to its ease, flexibility, and accuracy in the operation over distance [35]. Hence, we adopted IoT-based approach for the online monitoring. The main benefit of checking hostel room's status about tubes, lights, or fans at anytime from anywhere.

3 Proposed System Software and Hardware Architecture

The proposed system is built around Arduino version 1.6.9 for the ATmega328 and ESP8266. Arduino is the latest module, which is used for programming and compiling. For circuit design, we used OrCAD 9.2 version. For web application and for data management, we used PHP and HTML. IR transmitter and receiver are used for sensing the presence of student in the rooms.

3.1 Hardware Implementation

The modules used like ESP8266 Wi-Fi Module for on-chip wireless networking, the Atmega328P Atmel AVR Arduino microcontroller to optimize the power consumption versus processing speed, IR sensor where infrared radiation is smaller than microwaves but longer than visible light wavelengths that is 0.75–1000 μm and the optocoupler, also known as opto-isolator which is sensitive to light.

3.2 Software Implementation

Algorithmic steps:

1. Start
2. Initialized the IR MODULE, DIPLAY, ARDUINO, ESP8266.
3. Now, initialize the web page with the help of boot program.
4. According to entry or exit to the room, increment or decrement of the counter is done.
5. Running program for counter mode for each entry is IR 1, increment counter counts its 1.
6. Running program for counter mode for each exit is IR 2, decrement counter counts its 0.
7. Now, counter check whether it is "0" or not.
8. If it is not "0" then it returns to step 7.
9. If it is "0" then check status of optocoupler sensor for light ON or OFF.
10. Show this status to web page.
11. Now, check the switches is ON or OFF.
12. If yes, then make relay to turns OFF the switch. Otherwise, no action can be done.

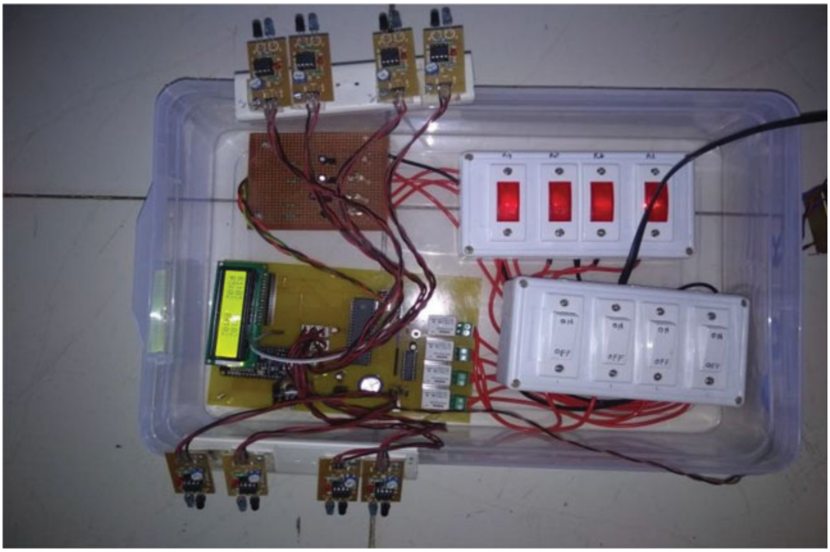


Fig. 1 Our implemented system

13. This information is saved on web page as an excel sheet.
14. When whole steps are done, then if light switch is ON and our system turns OFF, then system send a SMS for penalty of 100 Rs. to the students at their mobile number.

4 Experimental Result

The below Fig. 1 shows the implemented setup of our project work.

For online status check, here, is the website http://elpro.org.in/myproject/power_management/HTML/index.html. This website is first copied to the Google for checking room status. Figure 2 shows the webpage of our project. This is the main page, from where we login to the sites. Here, we need ID and PASSWORD.

After login and project's switch ON, it shows the default value of room status which is 2, after checking the entrance or exit from the room, Fig. 3 shows the room status on LCD display. Here, room No. 3 status is zero and it means room is empty.

As an example, Table 1 shows software implemented results of the online status of the rooms on the authorized person's personal computer, few comparative reading is also added related to similar work.

Table 1 The result displayed on the authorized person's PC

Sr. no.	Room No. 1 Count	Room No. 2 Count	Room No. 3 Count	Room No. 4 Count	Room No. 1 Status	Room No. 2 Status	Room No. 3 Status	Room No. 4 Status	Date/time
1	02	02	00	02	ON	ON	OFF	ON	2017-12-17 14:11:24
2	02	04	01	00	ON	ON	ON	OFF	2017-11-11 11:59:57
3	05	00	05	03	ON	OFF	ON	ON	2017-10-31 18:20:24

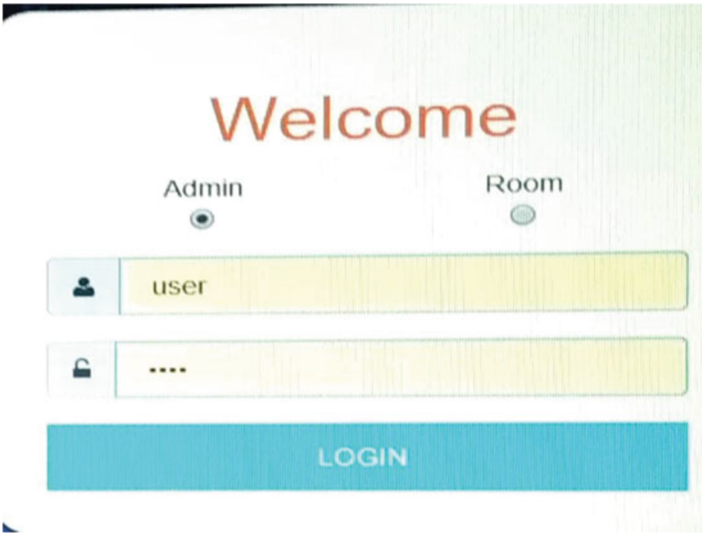


Fig. 2 Webpage of our system

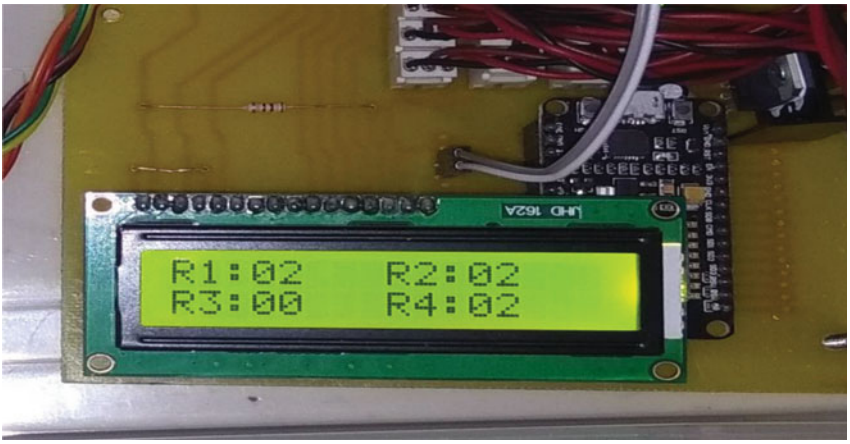


Fig. 3 R1, R2, R4 has two students and R3 is empty

Figure 4 shows the SMS to the students who leave the room without turning OFF the switches of fans and tubes lights. This process proceeds after the above steps are done.

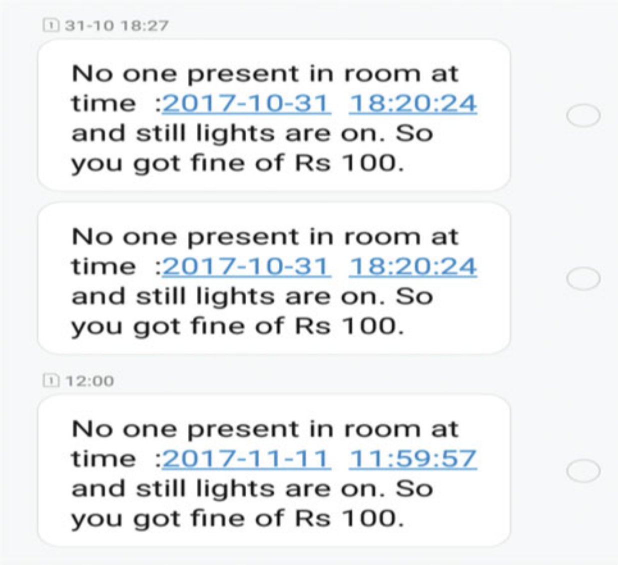


Fig. 4 SMS for fine to the student's mobile

5 Conclusion

We have presented the idea of online monitoring using Internet of Things. The main benefit from this project is that we can check the status of hostel rooms at anytime from anywhere only by Internet. The system can also send a SMS for penalty to the registered mobile number. The implemented system proves to be effective in energy conservation. This system is specially designed for power conservation. Some future scope to implement this system is supermarkets, airports, pharmaceutical companies, animal farms, libraries, hospitals, container yards, shipping, warehousing, car parking, etc.

References

1. Barreiro, C., Ferreira, B., Blanck, M., Abreu, C.: Energy management for rational electricity use (cement plant). In: XXXII, Record of Conference Papers, pp. 207–236. IEEE, 06 Aug 2002
2. Hassan, M.K., Mustafa, M., Jala, I.A.A.: Lighting management system. In: IEEE Student Conference on Research & Development—2003, SCORED 2003, pp. 309–312, 05 July 2005
3. Bell, D.G., Wils, T.L.: Energy conservation and demand control using distribution automation. In: IEEE Rural Electric Power Conference, 2004, C4-1-12, 28 June 2004
4. Zhang, C., Cui, N., Zhong, M., Cheng, Z.: Application of Fuzzy Decision in Lighting Control of Cities European Control Conference. In: 44th IEEE European Control Conference, CDC-ECC '05, pp. 4100–4104, 30 Jan 2006

5. Berenguer, M., Giordani, M., Giraud, F.: Automatic detection of activities of daily living from detecting and classifying electrical events on the residential power line. In: 10th IEEE International Conference on e-Health Networking, Application & Services, pp. 29–32, 15 Aug 2008
6. Han, J., Lee, H., Park, K.-R.: Remote-controllable and energy-saving room architecture based on ZigBee communication. In: IEEE International Conference on Consumer Electronics, ICCE'09, pp. 1–2, Feb 2009
7. Li, C., Xu, Q., Jiang, M., Chen, S.: Correspondence optical fiber automatic monitoring system. In: 9th IEEE International Conference on Electronic Measurement & Instruments, ICEMI'09, pp. 3-215–3-217, 02 Oct 2009
8. Singh, R.S.S., Ibrahim, A.F.T., Salim, S.I.M.D., Chiew, W.Y.: Door sensors for automatic light switching system. In: IEEE Third UK Sim European Symposium on Circuit Modeling & Simulation, EMS '09, pp. 574–578, 8 Dec 2009
9. Bin, S., Guiqing, Z., Lin, Z., Ming, W.: Multi-agent system design for room energy saving Industrial Electronics and Applications. In: 2010 the 5th IEEE Conference Industrial Electronics & Applications (ICIEA), pp. 73–76, 23 July 2010
10. Han, J., Choi, C.-S., Lee, I.: More efficient home energy management system based on ZigBee communication and infrared remote controls. In: IEEE International Conference on Consumer Electronics (ICCE '11), pp. 544–554, Feb 2011
11. Kadir, K.M., Forhad, Md.S., Fadlullah, M.M., Quader, N., Al-Arif, Md.M.R.: Energy saving by automatic control of power in simple home. In: 2011 IEEE 3rd International Conference on Communication Software & Networks (ICCSN), pp. 544–545, 08 Sept 2011
12. Martirano, L.: Lighting systems to save energy in educational classrooms. In: IEEE 10th International Conference on Environment & Electrical Engineering (EEEIC '11), pp. 1–5, 13 June 2011
13. Wei, C., Li, Y.: Design of energy consumption monitoring and energy saving management system of intelligent building based on the Internet of things Control. In: 2011 IEEE International Conference Electronics Communication & Control (ICECC), pp. 3650–3652, 03 Nov 2011
14. Veleva, S., Davcev, D., Kacarska, M.: Wireless smart platform for Home Energy Management System. In: 2nd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies Europe, pp. 1–8, 05 Mar 2012
15. Akpama, E.J., Okoro, O.I.: Economic implication of energy saving bulbs in the Cross-River University of Technology students' hostel. In: 20th IEEE International Conference on Domestic Use of Energy (DUE), pp. 79–84, 10 May 2012
16. Bjelica, M.Z., Mrazovac, B., Vojnovic, V., Papp, I.: Gateway device for energy—saving cloud—enabled smart home. In: 35th IEEE International Convention on MIPRO, pp. 865–868, 16 July 2012
17. Hsu, C.-L., Lin, S.-E.: Design for middle-way of intelligent energy-saving system with embedded system. In: Third IEEE International Conference Digital Manufacturing and Automation (ICDMA '12), pp. 319–322, 13 Sept 2012
18. Zanoli, S.M., Barchiesi, D.: Thermal and lighting control system with energy saving and users comfort features. In: 2012 20th Mediterranean Conference on Control & Automation (MED), pp. 472–477, 13 Aug 2012
19. van Rensburg, J.F., Mathews, E.H., Pelzer, R.: Energy management audit and recommendations at a tertiary institution. In: 9th IEEE International Conference on Industrial & Commercial Use of Energy (ICUE), pp. 1–6, 15 Oct 2012
20. Mangkalajan, S., Julsereewong, P., Julsereeong, A.: Building power management for energy saving: a case study of university building. In: 2012 12th IEEE International Conference on Control, Automation & System (ICCAS '12), pp. 1778–1782, 31 Dec 2012
21. Baraka, K., Ghobril, M., Malek, S., Kanj, R., Kayssi, A.: Low cost Arduino/Android-based energy-efficient home automation system with smart task scheduling. In: 5th IEEE International Conference on Computational Intelligence, Communication System & Networks (CICSyN), 2013, pp. 261–301, 01 Aug 2013

22. David, K., Klein, B.N.: A novel future Internet smart grid application for energy management in offices. In: IEEE International Conference on Future Network & Mobile Summit, 2013, 3–5 July 2013, pp. 1–10, 17 Oct 2013
23. Goyal, P., Shiva Kumar, B., Sudhakar, K.: Energy audit: a case study of energy centre and Hostel of MANIT. In: Bhopal 2013 International Conference on Green Computing, Communication & Conservation of Energy (ICGCE), pp. 644–648, 02 June 2014
24. Xu, Z., Jia, Q.-S., Guan, X., Xie, X.: A new method to solve large-scale building energy management for energy saving. In: 2014 IEEE International Conference on Automation Science & Engineering (CASE), pp. 940–945, 30 Oct 2014
25. Adriansyah, A., Dani, A.W.: Design of small smart home system based on Arduino. In: 2014, IEEE International Conference on Electrical Power, Communication, Control and Information Seminar (EECCISM ’14), 27–28 Aug 2014, pp. 121–125, 08 Jan 2015
26. Vergara, A.L., Villaruz, H.M.: Development of an Arduino-based automated household utility power monitoring system. In: 2014 International Conference on Humanoid, Nano technology, Information Technology, Communication & Control, Environment & Management (HNICEM), pp. 1–6, 22 Jan 2015
27. Patel, D., Khirsariya, A., Shah, J.: Automated system for hotel room service. In: 2014 2nd IEEE Emerging Technology Trends in Electronics, Communication & Networking (ET2ECN), pp. 1–6, 19 Feb 2015
28. En, Y.S., Fan, P., Ping, C.Z., Hua, S.W., Peng, Z.: Energy-saving system for classroom based on campus card. In: 2015 27th Chinese Control Conference (CCDC), pp. 5860–5864, 20 July 2015
29. Bhuvaneswari, S., Satish, B., Mahalaksmi, R.: Wireless Home Energy Consumption Control based on prioritized load switching. In: 2015, IEEE International Conference on Smart Technology & Management for Computing, Communication, Control, Energy and Materials (ICSTM), pp. 548–543, 27 Aug 2015
30. Méndez, G., Casillas, M.A., Baltazar, R., Lino, C., Mancilla, L., Lopez, S.: Intelligent management system for the conservation of energy. In: 2015 IEEE International Conference Intelligent Environment, pp. 120–123, 13 Aug 2015
31. Dey, A.N., Basak, P.: Simulation of a proposed PV based microgrid for a hostel building in Thapar. In: Michael Faraday IET International Summit 2015 Date of Conference on Michael Faraday IET, pp. 459–464, 28 July 2016
32. Mary Reena, K.E., Jacob, L., Mathew, A.T.: Energy efficient wireless networked Building Automation System controlled by real occupancy. In: 2015 IEEE region 10 International Conference (TENCON), pp. 1–6, 07 Jan 2016
33. Chaphekar, S.N., Mohite, R., Dharme, A.A.: Energy monitoring by energy audit and supply side management. In: 2015 IEEE International Conference on Energy Systems & Applications, pp. 178–183, 04 July 2016
34. Patil, M., Chaudhari Vijay, D., Dhande, H.V.: Hostel rooms power management and monitoring. In: Proceedings of National Conference on Current Trends in Engineering, Science, Technology and Management (NACCTESTM-2017), pp. 62–66, 21 Mar 2017. Published in: International Journal of Advances in Electronics and Communication System, approved by CSIR-NISCAIR
35. Deshpande, P., Iyer, B., Research directions in the Internet of Every Things(IoET). In: International Conference on Computing, Communication and Automation (ICCCA), pp. 1353–1357 (2017)

Performance Analysis of LAN, MAN, WAN, and WLAN Topologies for VoIP Services Using OPNET Modeler

Poonam Chakraborty and Aparna M. Telgote

Abstract Visual and Vocal communication can be transferred through Circuit switched Network or Packet Switched Network. Public Switched Telephone Network (PSTN) is not an affordable option therefore over existing packet switched network, Voice over Internet Protocol (VoIP) has become a preferable alternative due to its reduced cost. However, despite its reduced cost it has so many challenges which affect its successful deployment. This is because; the quality of VoIP is mainly affected by jitter, delay, packet loss and some other parameters. This research was carried out to evaluate voice quality in VoIP experimentally under different scenarios using OPNET network simulator. A VoIP network was simulated using Riverbed modeler academic edition 17.5 and the behavior and quality of VoIP was studied and analyzed under different scenarios. The results of the analysis and the performance evaluation are presented in this paper. This work can guide researchers and designers to design a network for VoIP services and its deployment. It can also guide the operators to choose speech compression technique for better voice quality.

Keywords Voice over Internet Protocol (VoIP) · Quality of Service (QoS) Mean Opinion Score (MOS) · Jitter

1 Introduction

With the advancements in telecommunication technologies there evolved the concept of converged network. The converged networks uses technology that in turn uses Internet as a medium to transmit data, voice and video information in the form of packets uses packet switching. It has numerous benefits by providing add-on to both the service providers and users [1]. The VoIP technology converts voice

P. Chakraborty (✉) · A. M. Telgote
Don Bosco Institute of Technology, Mumbai, India
e-mail: poonam@dbt.in

A. M. Telgote
e-mail: aparna@dbit.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_20

Table 1 Mean Opinion Score (MOS)

Score	Quality	Scale of listening effort
5	Excellent	No effort is required
4	Good	No considerable effort is required
3	Fair	Moderate effort is required
2	Poor	Considerable effort is required

sample to binary code, which is then stored and transferred in the form of packets. These packets are then transferred together with thousands of other different or same data packets on some existing network transmission line [2]. In case of circuit switching in PSTN, a dedicated channel is reserved over the whole duration of a call, whereas with the packet switched internet, bandwidth is consumed only when voice packets are delivered. But Voice over Internet Protocol (VoIP) technology improves the bandwidth and facilitates creation of new varied services. In packet switched network data traffic delays are acceptable in the data traffic flow or the system is asynchronous and extremely sensitive to errors, while in case of voice traffic significant delays are not acceptable or the system is synchronous and more tolerant of errors [3].

If we intend to make the system more efficient and also to provide better quality of service, VoIP is the solution in any business or organization [4] that helps in adding more features, cuts down operational costs and increases the interaction between employees and customers [5]. Though VoIP has many advantages such as reduced cost, accessibility, etc., but it has some challenges such as security and QoS [6, 7].

VoIP security [8] should not be overlooked by the users as this technology has some inherent security risks and threats associated with it that can affect the organization's confidentiality, integrity, and availability. This security risk result from its architecture which is different from that of the traditional circuit switched based telephone network [9].

2 QoS Parameters of VoIP Traffic

In VoIP, the major factors that affect Mean Opinion Score (MOS) which is measure of voice are jitter, echo, speech compression, and packet loss. It is based on a perceptual scale of 1–5 as shown in Table 1.

The factors like jitter, echo, speech compression and packet loss that affect MOS are discussed below [10].

Table 2 Comparison of CODEC

CODEC	Bit rate in kbps	Ie	MOS
G.711	64	0	4
G.726	32	7	3.85
G.728	16	7	3.6
G.729	8	11	3.7
G.723.1	6.3	15	3.6
G.723.1	5.3	19	3.9

2.1 Jitter

It is defined as the latency in the variation in arrival rate of voice packets at the receiver/destination. The main effect of jitter control is its impact on latency and packet loss. The packet loss increases if the jitter size exceeds the size of the buffer, and the VoIP packet is discarded. In order to minimize the packet loss if the jitter buffer size is increased then the end-to-end latency is increased as more time is spent in the buffer. This results in a degradation of performance due to high latency. Therefore, the requirement is that there should be a trade-off between latency and packet loss.

2.2 Echo

Sometimes the speakers own voice leaks in their own receiver path. This is called echo—the audible leak-through of a speaker's voice into their own receiver path. The "leak through" occurs at the end of the instrument used or in an analog circuit instrument. Sometimes there is poor insulation between transmitter and receiver cables and it can also cause leak-through.

2.3 Speech Compression

It uses predictive coding that reduces the bit rate required for transmission of the voice carrying stream. Impairment factor (Ie) is considered the impact caused by speech compression due to distortion of the voice during the compression technique. Comparison of Codec is presented in Table 2 [11].

Table 3 ITU-T specification for delay and jitter

Network parameters	Good	Acceptable	Poor
Delay (ms)	0–150	150–300	>300
Jitter (ms)	0–20	20–50	>50

2.4 Packet Loss

It defines success rate of packet transmission and the measures is given in terms of percentage. The ITU-T guidelines for the delay and jitter for the different types of call quality, as listed in Table 3.

3 Methodology

Many challenges are faced by designer of VoIP. This research was carried out to evaluate voice quality in VoIP experimentally under different scenarios using OPNET network simulator. A VoIP network was simulated using Riverbed modeler academic edition 17.5 [12] and the behavior and quality of VoIP was studied and analyzed under different scenarios. In this research paper the focus is on QoS [13].

The aim of this research [14, 15] is to investigate the effect of the following factors on system performance:

- Increasing the number of VoIP clients
- Traffic arrival distributions
- Voice codec schemes
- Conference call
- Effect of Interference

4 Implementation Scenarios and Results

4.1 Scenario-1 Analysis of Parameters with Respect to Number of Nodes in LAN

Initially only two clients are taken and simulation done. Later the number of nodes/clients increased to twenty and fifty respectively. The network has star topology. The components used are Ethernet switch, router, Ethernet workstations and 10 base-T Duplex links as shown in Fig. 1.

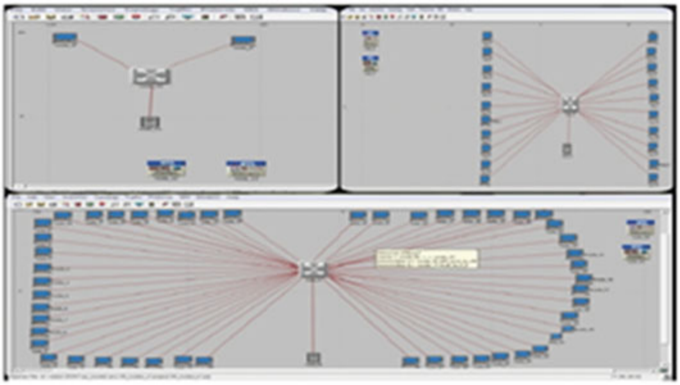


Fig. 1 LAN network model with 2, 20 and 50 nodes

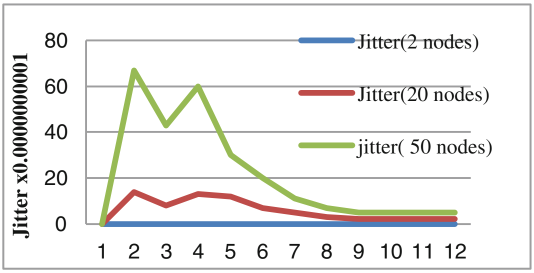


Fig. 2 Jitter of the simple LAN network model

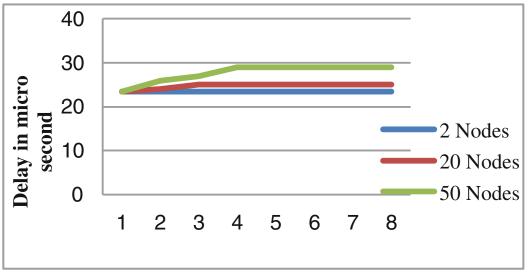


Fig. 3 Delay of the simple LAN network model

The Jitter, delay and MOS outputs obtained are shown in Figs. 2 and 3 respectively. The blue, red and green line indicates two, twenty and fifty nodes respectively. Note in all the graphs the x-axis represents the time in milli seconds.

Figure 3 shows that a slight amount of delay always occurs. Delay is affected by the total number of clients placing a VoIP call. The increase in delay with the number of clients is quiet expected because the packets are likely to take more time so as to select the correct receiver client out of all the clients.

Table 4 Result of analysis of VoIP parameters with respect to number of nodes in LAN

Nodes	Delay	Jitter	MOS	Packet delay variation	Packet end-to-end delay	Traffic sent	Traffic received
2	Remains constant at 0.236 ms	0	Remains constant at 4.3612	Delay in packets observed from 110 s	Delay increases at 145 s	Low	Same as Traffic sent
20	Slight variation at start then constant at 0.25 ms	Peak is observed with maximum value of 9.8 ns	Starts from 4.36121 but settles down at 4.36115	Delay in packets observed from 101 s	Delay increases at 114 s	Medium	Same as Traffic sent
50	Noticeable variation at start then constant at 0.292 ms	Peak is observed with maximum value of 33 ns	Starts from 4.36121 but settles down at 4.36104	Delay in packets observed from 100 s	Delay increases at 100 s	High	Same as Traffic sent

The MOS value decreases with increase in the number of VoIP clients. This is because voice quality in VoIP is basically bandwidth dependent. As the number of client increases the bandwidth available also decreases and hence the quality of voice and MOS decreases. The various other simulation results of the analysis of VoIP with respect to the number of nodes in LAN is shown in Table 4 below.

From the result summarized in Table 4 we can observe that the quality of VoIP call deteriorates as the number of clients in a VoIP network increases. When the VoIP clients increases, overload happens and causes larger fluctuation in jitter, longer end-to-end delay, lower MOS value and more packet loss.

4.2 Scenario 2: Local VoIP Versus Long Distance VoIP

In long distance VoIP network two subnets are added on the map and each represents a company or office. The cloud symbol represents the Internet. In Local VoIP Network LANs, it has been modeled as subnets. Each LAN contains an Ethernet switch, a Cisco router and various workstations. The configuration used here is star as shown in Fig. 4.

One way to assign the VoIP application to our model can be made under the Application Definition. Application configuration files contain settings specific to an application. This file contains configuration settings that the common language

Fig. 4 Local and long distance VoIP call model in LAN

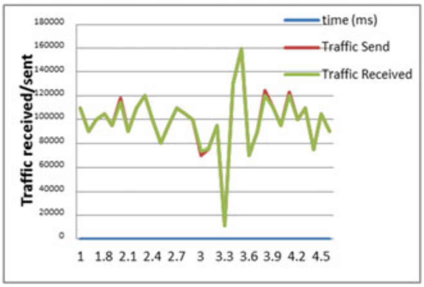


Fig. 5 Traffic sent and received observed in local VoIP call model

runtime reads and settings that the application can read. Application configuration and profile configuration are used together to generate the traffic.

After simulation we observe that the traffic received is less than the traffic sent in Long Distance VoIP call model. On the other hand the Traffic Sent and Received in the Local VoIP call model is exactly the same. Hence a local VoIP call has no packet loss even if some small amount of delay is present. But on the other hand a Long Distance VoIP call does suffer some amount of packet loss since the traffic sent and received is not the same. The red line represents traffic sent and the blue line represents traffic received as shown in Figs. 5 and 6.

The summary of various simulation parameters are listed in Table 5.

Fig. 6 Traffic sent and received in the long distance call model

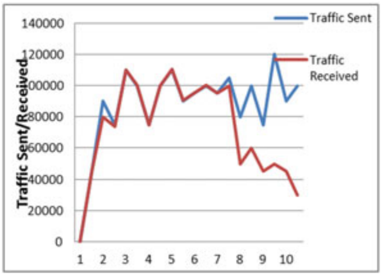


Table 5 Result of local VoIP and long distance VoIP calling

Scenario	Delay	Jitter	MOS	Packet delay variation	Packet end to end delay	Traffic sent	Traffic received
Local VoIP	Maximum delay of 0.177 ms	Very negligible almost 0 s	Value of 1.1 through-out	Remains more or less constant at 0.118 μs	Constant at 0.1 s	Low as compared to Long distance VoIP	Same as Traffic sent
Long distance VoIP	Delay increases with maximum of 1.4 s	Jitter is increasing with time. Maximum value is 0.556 ms	Maximum value of 3.29 then drops to 1.1 eventually	Increases with time. Maximum value of 7.00 s	Increase with time. Maximum value of 9.45 s	High as compared to Long distance VoIP	Is less than Traffic sent

4.3 Scenario 3: Ethernet Versus WLAN

Initially a scenario of by WLAN and LAN network with 20 nodes is constructed. The quality of the VoIP call in both the approach is compared in a LAN or WLAN network as shown in Fig. 7.

The ITU standards specify that the jitter should be less than 75 ms but not greater than 40 ms. Jitter is almost 0 in case of a LAN network, however jitter increases with time.

The ITU standards specify that the jitter should be less than 75 ms but not greater than 40 ms. Jitter is almost 0 in case of a LAN network, however jitter increases with time and the peak value of approximately 2.2 ms is observed in case of WLAN [16]. Jitters show tremendous variations in a WLAN network as compared to a LAN network. The various other parameters were simulated for the LAN and WLAN setup. A comparative analysis of Ethernet and WLAN is given in table number 6. From Table 6, we can conclude that Ethernet is better as compared to WLAN when it comes to variations in factors that influence performance of a VoIP call. High cabling costs, more infrastructure size and immobile devices make it difficult for an Ethernet model to be implemented in real life.

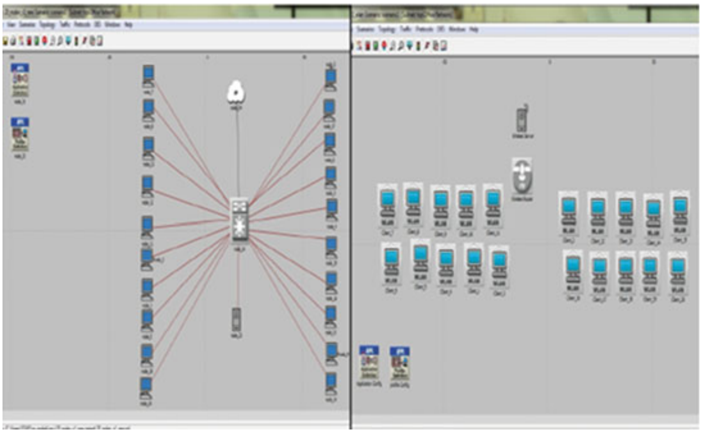


Fig. 7 LAN and WLAN model with 20 nodes

Table 6 Comparison of Ethernet versus LAN

Parameters	Ethernet	WLAN
Jitter	Low	High
MOS	High	Low
Packet end to end delay	Low	High
Packet delay variation	Low	High
Traffic sent	Same as WLAN	Same as LAN
Traffic received	Same as traffic sent	Traffic is lost

4.4 Scenario 4: Audio Codec Schemes

Here two different encoder schemes G.711 and G.729 are used as shown in Fig. 8. In both the cases silence suppression is enabled G.729 is the one that offers high quality as well as robust speech performance quality at the price of increased complexity. The G.729 speech coder is an 8 kbps Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP) speech compression algorithm approved by ITU-T [17, 18]. G.729 requires 10 ms input frames and on the other hand generates frames of 80 bits in length. With the vocoder processing signals of 10 ms frames and 5 ms look-ahead, the total G.729 algorithmic delay comes out to be 15 ms.

In the profile configuration we select the various codec schemes for example G.711 as shown in Fig. 9. The simulation results obtained for various parameters using audio codec schemes G.711 and G.729 are listed in Table 7.

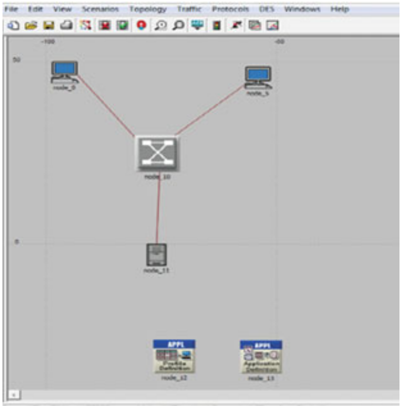


Fig. 8 Model for audio codec schemes

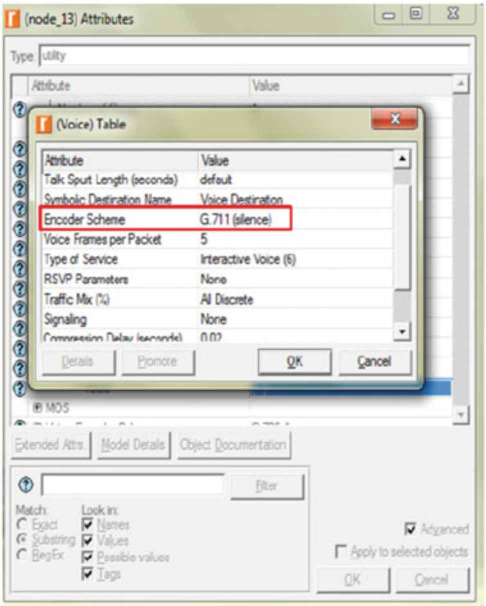


Fig. 9 Application configuration in the G.711 network model

5 Conclusion

The research aimed to increase the QoS in VoIP. There are various approaches for simulations like ns2, Wandl, Opnet Modeler, etc. The use of Opnet Modeler helped us to simulate the various scenarios as it is open free software (academic version) and uses simple modeling techniques. After implementing in Opnet modeler various scenarios, it was found that all the parameters like delay, jitter, packet delay variation, packet end-to-end delay, MOS, traffic sent and traffic received increases with the increase in the number of nodes (clients). With the increase in the LAN distance

Table 7 Result of Audio Codec Scheme G.711 and G.729

Scenario	Delay	Jitter	MOS	Packet delay variation	Packet end to end delay	Traffic sent	Traffic received
G.711	Less. Maximum value of 0.188 ms	0 s	Almost constant value of 1	Very less almost in μs	Constant at 0.19 s	Very Low	Same as Traffic sent
G.729	High. Maximum value of 0.5 s	Starts with jitter of 0 s then rises to 0.5 ms	Value of 4	As high as 1 s	Increases to 3.7 s	Very High	Is less than Traffic sent

there is no improvement in the output values, rather the parameter values increases and so the Quality Service (QoS) deteriorates. In the wireless scenario it was observed that the output parameters increase with the increase in the number of nodes (clients). To optimize the network encoder scheme G.711 and G.729 was implemented and it was found that by using G.711, there was an increase in the bandwidth which is good for any communication system. So we can conclude that by using G.711 encoder the bandwidth efficiency increases tremendously. Also in future we can implement various other sophisticated methods to improve the QoS.

References

1. Ayokunle, O.O.: Integrating Voice over Internet Protocol (VoIP) technology as a communication tool on a converged network in Nigeria. ICT J. **2**(11) (2012). ISSN 2223-4985
2. Sze, H.P., Liew, S.C., Senior Member IEEE, Lee, J.Y.B., Yip, D.C.S.: A multiplexing scheme for H.323 voice-over-IP applications. IEEE J. Sel. Areas Commun. **20**(7) (2002)
3. Lee, C.-Y., Kim, H.-K., Ko, K.-H., Kim, J.-W., Jeong, H.-C.: A VoIP traffic monitoring system based on Net Flow v9. Int. J. Adv. Sci. Technol. **4** (2009)
4. Kos, A., Klepec, B., Tomazic, S.: Technique for performance improvement of VoIP applications. In: Electro Technical Conference, MELECON 2002, Published in IEEE Xplore (2002)
5. Uys, L.: Voice over internet protocol (VoIP) as a Communications tool in South African business. Afr. J. Bus. Manag. **3**, 089–094 (2009). ISSN 1993-8233
6. Lee, J., Kim, S.: VoIP-aware network attack detection based on statistics and behaviour of SIP traffic. Springer link, pp. 001–009 (2014)
7. Ayokunle, O.O.: Implementing security on a Voice over Internet Protocol (VoIP) network: a practical approach. IOSR J. Comput. Eng. (IOSRJCE) **7**(4) (2012). ISSN: 2278-0661, ISBN: 2278-8727
8. Sonkar, S.K., Singh, R., Chauhan, R., Singh, A.P.: A review paper: security on voice over internet protocol from spoofing attacks. Int. J. Adv. Res. Comput. Commun. Eng. **1**(3) (2012)
9. Shah, K., Ghrera, S.P., Thaker, A.: A novel approach for security issues in VOIP networks in virtualization with IVR. Int. J. Distrib. Parallel Syst. (IJDPS) **3**(3) (2012)
10. Al-Sayyed, R., Pattinson, C., Dacre, T.: VoIP and database traffic co-existence over IEEE 802.11b WLAN with redundancy. World Acad. Sci. Eng. Technol. Int. J. Electr. Comput. Electron. Commun. Eng. **1**(2) (2007)

11. Alsahlany, A.M.: Performance analysis of VoIP Traffic over integrating wireless LAN and WAN using different codec. Int. J. Wirel. Mob. Netw. **6**(3), 79–89 (2014)
12. Riverbed Modeler: The fastest discrete event-simulation engine for analyzing and designing communication networks. https://support.riverbed.com/content/support/software/opnetmodel/modeler.html. Accessed 2 Dec 2017
13. Wang, S., Mai, Z., Xuan, D., Zhao, W.: Design and implementation of QoS-provisioning system for voice over IP. IEEE Trans. Parallel Distrib. Syst. **17**(3) (2006)
14. Yang, S., Zhang, J.: New idea to simulate performance of VoIP network with OPNET. In: International Conference on Internet Technology and Applications, 2010, published in IEEE Xplore (2010)
15. Lili, C.: VoIP system simulation design and implementation. In: Second International Conference on Instrumentation, Measurement, Computer, Communication and Control (IMCCC), added to IEEE Xplore (2013)
16. Atif Qureshi, M., Younus, A., Saeed, M., Sidiqui, F.A., Touheed, N., Shahid Qureshi, M.: Comparative study of VoIP over WiMax and Wi-Fi. IJCSI Int. J. Comput. Sci. Issues **8**(3), 1 (2011). ISSN (Online): 1694-0814
17. Shum, H.-Y., Liao, M., Chang, S.-F.: Codec schemes selection for wireless voice over IP (VoIP). In: Pacific Rim Conference on Multimedia PCM 2001, LNCS 2195, pp. 622–629. Springer, Berlin, Heidelberg (2001)
18. Ismail, M.N.: Analyzing of MOS and Codec Selection for Voice over IP Technology, Anale. Seria Informatica. Vol. VII fasc. 1– 2009 Annals. Computer Science Series. 7th Tome 1st Fasc., pp. 263–276 (2009)

Intelligent Attribute Based Encryption (IABE) Mechanism for Health Records in Cloud

Ranjith Kumar Vollala and L. Venkateswara Reddy

Abstract Various possible definitions are to be found in appropriated figuring. A vast bit of them focuses on the development in a manner of speaking. Research has been done to merge all these particular definitions to come up with one uniform definition. Conveyed processing can best be depicted as a mammoth pool which contains gear, programming and diverse organizations that can be gotten to through the "cloud". Each one of these advantages can be gotten to at whatever point generally. A significant part of the time the provider of the cloud offers his organization as pay-per-use. This infers there is high versatility in the use of these organizations as extra resources are always available. Moving fragile data from in-house IT system to a cloud arrange has transformed into a mind-boggling and testing undertaking. This paper proposes another protected technique to secure customer fragile data. In this methodology, we are considering therapeutic administrations data. We are using multilevel quality based encryption plot for securing customer's wellbeing records. The trial comes to fruition demonstrate ideal results over existing techniques.

Keywords Encryption · Health care · Cloud · PHR · Key

1 Introduction

Public Health Record (PHR) has been created as a patient-driven model of the strategic of well-being data by utilizing the cloud innovation. PHR could be viewed as the answer for the better administration of a person's well-being, and as the instrument that will enable the patient in connection with social insurance suppliers through the capacity to give his/her own particular therapeutic history. Moreover, PHR enables

R. K. Vollala (✉)
Department of CSE, Rayalaseema University, Kurnool, Andhra Pradesh, India
e-mail: ranjithphdd@gmail.com

L. Venkateswara Reddy
Department of IT, Sree Vidyanikethan Engineering College, Chittoor, Andhra Pradesh, India
e-mail: lakkireddy.v@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_21

a patient to make, recover and deal with his/her own well-being information record starting with one place then onto the next place, through a cloud server. For instance, Alice may first observe a specialist at Clinic A. At that point the specialist alludes her to Specialist B. Alice may likewise need to take the blood test at research facility C. At last Alice goes to Hospital D for obtaining restorative treatment. By utilizing PHR, Alice does not have to bring along or exchange his/her paper-based therapeutic record among various medicinal services suppliers, which adds to the change of patient's solace level by utilizing cloud innovation. In medicinal services space, the quickest developing region is m-Health or portable human services. It means that entrance to medicinal services information or electronic wellbeing records of patients is given through cell phones. This has turned into a reality because of the approach of new advances and accessibility of administrations of 3G and 4G [1]. As individuals of all kinds of different backgrounds are utilizing cell phones and a little hand held gadgets, the utilization of electronic social insurance records should be possible through cell phones [2]. In the human services space, the patient care is changing from individual contact with specialists to different means. Remote well-being observing frameworks, telemedicine is some new ideas that are getting to be a noticeably main stream as of late. There is proof of versatility and interoperability of wellbeing administrations. The most recent systems administration innovations like Long-Term Evolution (LTE), 4G and WiMAX guarantee better approach for rendering data to clients. These advancements empower hand held gadgets likewise to make utilization of media therapeutic information [3].

Tragically creating and sending a portable social insurance framework includes time, cash, and register assets. The issues here incorporate confinement to get to human services information, earlier ventures on inheritance frameworks, stockpiling and preparing necessities and the absence of dependable unified storehouse. A doctor's facility is a medicinal services association giving patient treatment by master doctors, specialists and types of gear. A report from a social insurance accreditation amass says that miscommunication amongst patients and human services suppliers is the purpose behind the hole in giving crisis therapeutic care to individuals in require. In creating nations, ignorance is the significant key root for passing's coming about because of dubious infections constituting a genuine general medical issue. Rationally influenced, diversely capable and oblivious patients cannot convey about their medicinal history to the therapeutic professionals. Human services are encountering an emotional change in data trade since the most recent decade. The definition edges of circulated figuring are very gone head to head in regards to today. Disseminated registering incorporates the sharing or limit by customers of their own information on remote servers had or worked by others and gets to through the Internet or distinctive affiliations. Appropriated registering organizations exist in various assortments, including data amassing goals, video areas; evaluate game plan goals, singular health record locales and some more.

2 Related Works

Security and protection in the social insurance distributed computing are something other than client benefits and secret word authorization. Distributed computing stages are multi-area conditions in which every space can utilize diverse security, protection, approaches and methodology, and trust prerequisites and possibly utilize different components, interfaces, and semantics, secure information reinforcement procedure, outsider confirmation. Such areas could speak to exclusively empowered administrations or other infrastructural or application segments. In human services cloud, security ought to be the best need from the very first moment. In restorative administrations cloud applications, a bit of the security and insurance issue and essentials are orthogonal to the strong cloud advantage model or cloud association indicate used. Around there, we rapidly demonstrate these issue and necessities. Security in Cloud Computing involves developed security plans, for instance, encryption, and get to organization, firewalls and interference disclosure. In inside Clouds handling the IT office can present all open security game plans which looks well yet in outside security of cloud computing depends upon the service provider of cloud (CSP). Few providers do not give in the option of safety game plans, as others let the utilization of customer safety necessities.

Part based access control (RBAC) [1] has been utilized to build up a broadened rendition, Open Architecture for Secure Interworking Services (OASIS) [2], to give a more adaptable access control system. Desert garden was utilized as a part of [3] to offer a hierarchical security administration framework to oversee get to control for approved faculty. The entrance control instrument acquainted approving different human services suppliers with seeing Electronic Health Records, along these lines sharing required EHRs. Then again, RBAC is deficient to deal with the adaptability expected to share EHRs due to their various basic and semantic nature. The patient-driven approval structure proposed in [4] considers totaling EHRs from various social insurance suppliers into virtual composite EHRs, at that point utilizes an entrance control conspire for composite EHRs [5] to give access to just approved clients. This work accepts patients utilize e-Consents [6] to control the sharing of their EHRs. E-Consents empower patients to give or withhold access to their electronic wellbeing data. A Unified Data Schema (UDS) reference display, as openEHR [7] and HL7, was produced to characterize nonspecific semantics and legitimate connections between quiet data components in EHR cases. The UDS permits a virtual piece of EHRs with a specific end goal to be collected. It likewise characterizes three properties (birthplace, affectability, and protest sort) for every hub that will be utilized to keep patients' delicate therapeutic data from being accidentally uncovered. Upon the patient's demand and e-Consent, particularly asked for EHR cases for a patient from various clinics are accumulated utilizing the UDS progressive structure to frame a virtual composite EHR. An entrance control approach is then created to recognize the approval impact on the collected virtual composite EHRs. Authors [8–10] discussed several threats and solutions regarding cloud data. Makers have developed concentrated piece of the assurance setting and security point of view in the e-cloud appropriated registering.

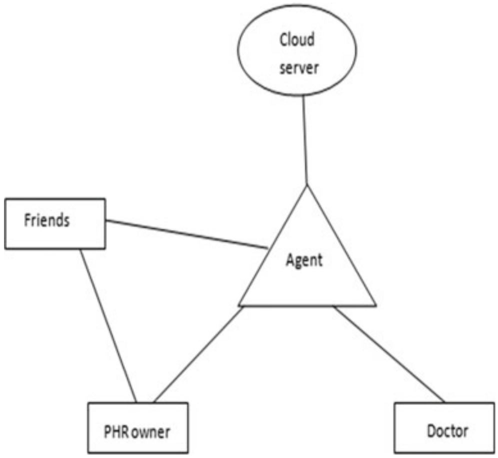
The key areas that impact the specific parts of security and assurance in social protection fogs are: data accumulating and getting ready, organization of e-cloud Infrastructure, convenience and customer experience. A huge bit of these issues can be overpowered by the providers of appropriated figuring organizations, frameworks of help of the e-human administrations system and customers themselves. The legal points suggest the security order and headings in the countries where use dispersed figuring. Courses of action on the arrangement of security sanctioning in the European Union and the United States are shifting. Assurance in the US is scattered among various assorted territory specific laws and these portions join the social protection fragment for the HIPAA. A few methodologies and arrangements have been proposed or executed. For example, the Netherlands executed a concentrated national EHR framework [11]. The framework utilizes a detached privacy assurance system: control and caution [12]. In this procedure, a checking framework logs all get to an EHR, given that all social insurance suppliers can get to every single patient datum, while the notice framework assesses the consistence of each entrance with getting to approaches that execute the law. It is discovered that, by and large, unlawful gets to are around 9% of all solicitations by medicinal services suppliers [12]. Creators [13] proposed a conveyed approach for ensuring the classification of EHRs. The principle thought is to epitomize the EHR information of every patient alongside the entrance approaches into a substance called dynamic package [14] and to store the package in the vault of the essential medicinal services supplier of the patient. Notwithstanding consolation from the government few state bureaus of adjustments have actualized electronic therapeutic records frameworks This parallels the free world where paper curative records keep on predominating [15]. There are numerous applications, for example, SONY's Lifelog's, Google Fit, and open m-Health accessible in the market, which measure day by day physical exercises, for example, strolling, running, bicycling, and transportation. They can likewise quantify diverse wellbeing related parameters, for example, heart rate, calorie level, and glucose level. The use case of the examples comprises web applications, building servers, [16].

3 Proposed Work

We display another structure as appeared in Fig. 1 called personal health record Agent. Here Agent is a line involving the Server of cloud and users. Here each client is going to upload his health record and particular key is given to agent. After that agent will generate control key from the user key. Then both keys are used for encryption kept in the appropriated stockpiling for the authenticity time allotment. After the authenticity time span, the Ck and its looking at information key is detached as of the dispersed stockpiling.
Algorithm for encryption

Input: PHR Key
Output: Cipher text

Fig. 1 Proposed model



1. Start
2. Select any integer ‘s’
Ck = m.es
Cipher Text CT = m mod Ck

Inside the legitimate era, if any entrance is asked for by the Ck the information key for encrypting text by unscrambling with the Ck. Using the information key records are decoded and given to client.

Algorithm for Decryption

Input: Control key
Output: Plain text
1. Select Ck
2. Decryption text=

Once the authority time is over, the Ck is removed from server and also consequent phr key is also removed. Regardless of the possibility that the scrambled substance is burglarized, with the control key alone client can’t decode the substance, on the grounds that the substance is encoded with information key. This guarantees time-based repudiation of access ideal for the clients of the information in the cloud.

4 Results

We consider Table 1 configuration for testing our method.

As shown in Fig. 2 the encryption time for the proposed system is giving a better result than the existing system. Here random key generation is faster so it will take less time for encryption. There are 100 attributes taken. We have ranged attributes

Table 1 System configuration

Cpu	Intel core i3 processor
Ram	8 GB
Hard disk	2 TB
OS	Windows 10
No of Nodes	8

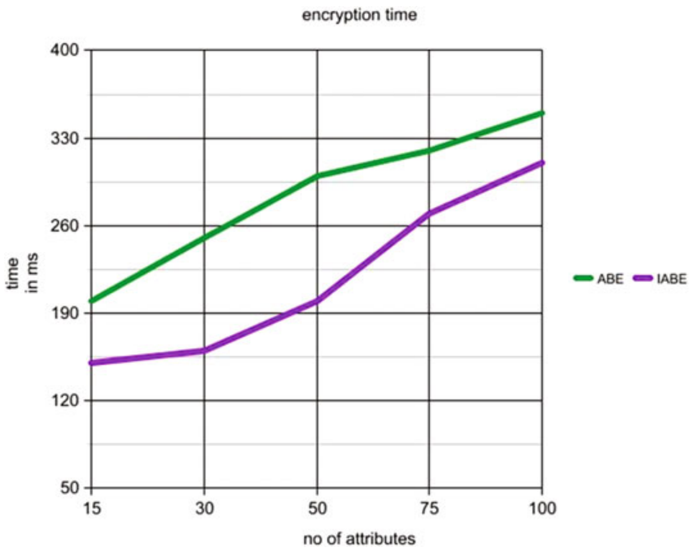


Fig. 2 Comparison of encryption time

from 15 to 100 in X axis. We calculated the time in milliseconds that is shown in y axis. For encrypting the 30 attributes existing ABE takes 250 ms whereas the proposed IABE takes only 150 ms. This shows the efficiency of the proposed system.

As shown in Fig. 3 the Decryption time for the proposed system is taking higher time than the existing system. Here number of iterations is more that is why decryption time is higher than existing one. As described in the Fig. 2 for encryption here also we conducted experiment for decryption time also. The proposed IABE decryption algorithm takes more time than existing ABE for decrypting attributes. It is because of the no of iterative steps involved in decryption.

5 Conclusion

The explained work is for securing personal records of health and sharing in cloud. Allowing for to some degree tried and true servers of cloud, we battle that to totally comprehend the client-driven thought, clients ought to have total manage of their hold

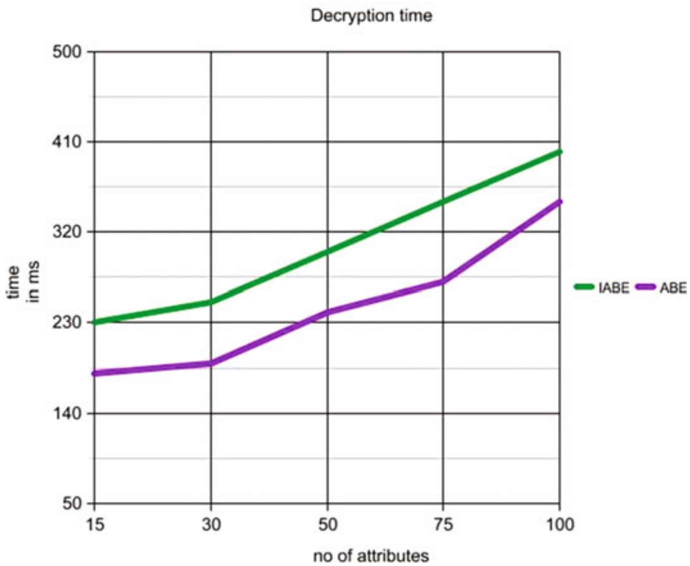


Fig. 3 Comparison of decryption time

security during scrambling their health reports to let to know there data. The structure watches out for the stand-out issues talked by various health record proprietors and customers, in that we colossally diminish the versatile nature of key organization while enhancing the security guarantees differentiated and past works. Here use of IABE to scramble data of health records, so user pat be able to let get to by singular customers, and additionally unique customers from open territories with different master parts, capacities, and affiliations. In addition, we overhaul an IABE plan to manage beneficial and on-ask for customer renouncement and exhibit its security.

References

1. Sandhu, R., Coyne, E., Feinstein, H., Youman, C.: Role-based access control models. IEEE Comput. **29**, 38–47 (1996)
2. Hine, J.H., Yao, W., Bacon, J., Moody, K.: An architecture for distributed OASIS services. In: Proceedings of International Conference on Distributed Systems Platforms and Open Distributed Processing, vol. 1795, pp. 104–120. Springer, Apr 2000
3. Eyers, D., Bacon, J., Moody, K.: OASIS role-based access control for electronic health records. In: IEEE Proceedings—Software, vol. 153, pp. 16–23 (2006)
4. Jin, J., Ahn, G.-J., Hu, H., Covington, M.J., Zhang, X.: Patient-centric authorization framework for sharing electronic health records. In: Proceedings of 14th ACM Symposium on Access Control Models and Technologies, pp. 125–134. ACM (2009)
5. Jin, J., Ahn, G.-J., Covington, M.J., Zhang, X.: Toward an access control model for sharing composite electronic health records. In: Proceedings of 4th International Conference on Collaborative Computing, Nov 2008

6. Coiera, E., Clarke, R.: e-Consent: the design and implementation of consumer consent mechanisms in an electronic environment. J. Am. Med. Inform. Assoc. **11**(2), 129–140 (2004)
7. openEHR Foundation: openEHR: an open domain-driven platform for developing flexible e-health systems
8. Reddy, L.V., Kumar, N.S.: Multi layer encryption using access control in public clouds. i-Managers' J. Cloud Comput. **2**(1), 24–31 (2015)
9. Ranjith, Venkateswara Reddy, L.: Threats –Solutions in Cloud security, Recent and Innovation Trends in Computing and Communication (IJRITCC)
10. Ranjith, Venkateswara Reddy, L.: An improved cryptographic mechanism for cloud storage system. Int. J. Appl. Eng. Res. (Accepted for its publication)
11. Perry Groot, M.O., Bruijsten, F.: Patient data confidentiality issues of the dutch electronic health care record. In: Proceedings of the 19th Belgian-Dutch Conference on Artificial Intelligence (BNAIC), pp. 151–157 (2007)
12. Salih, R.M., Lilien, L., Othmane, L.B.: Protecting patients electronic health records using enhanced active bundles. In: Proceedings of the 6th International Conference on Pervasive Computing Technologies for Healthcare, Doctoral Consortium, pp. 1–4 (2012)
13. Ben Othmane, L.: Active bundles for protecting confidentiality of sensitive data throughout their lifecycle. Ph.D. dissertation, Kalamazoo, MI, USA, aAI3441001 (2010)
14. Ranchal, R.: Cross-domain data dissemination and policy enforcement. Ph.D. dissertation, West Lafayette, IN, USA (2015)
15. Moore, P.L.: Taming the beast. 2006/2007 physicians practice technology guide. Physicians Practice, pp. 7–10 (2006)
16. Amazon, E.C.: Amazon Web Services EC2 instance types. http://aws.amazon.com/ec2/instance-types/ (2014). Accessed 10 Feb 2017, Accessed 21 Nov 2016

Latent Class Analysis (LCA) Based Approach for Finding Best Hotels

Vijay Singh, Bhasker Pant, D. P. Singh and Santosh Kumar

Abstract Researchers shows a huge interest in Latent Class Analysis (LCA) in various domains over the last two decades. We proposed a new Latent Class Data Analysis using Statistical modeling approach to categorize better and worst Hotel to Stay. The main objective of this study was to find the unobserved classes in the Trap Advisor dataset. The results allow to identify new entry of the Hotel and detects whether it lies in Good Hotel category or in worst Hotel category. For evaluation and demonstration purpose freely, available Trip Advisor dataset is used.

Keywords Latent Class Analysis (LCA) · Statistical modeling · Trip advisor

1 Introduction

As the number of hotels and tourists is increasing so rapidly, it is difficult to estimate whether the hotel is good to stay or bad. Many travel websites like Trip Advisor (https://www.tripadvisor.in/) or Orbitz (https://www.orbitz.com/) provide the recommendations to users based on some previous history or personalized data. As the data size increases day by day, there is an open challenge for everyone to extract relevant information from the existing dataset. Latent Class Analysis (LCA) is one of the ways to detect the unobserved group from the observed data. LCA is like conventional clustering techniques, but the difference is that it works on log-likelihood while others work on distance measure. LCA can be applied with dichotomous as well as with polytomous variable. LCA is somewhat like factor analysis, and it is used to classify groups based on maximum likelihood membership. Nevin L.Zhang [1] proposed a Hierarchical class model for cluster analysis; his major contribution was to develop an algorithm for learning Hierarchical latent class model from data. Although LCA

V. Singh · B. Pant · D. P. Singh · S. Kumar (✉)
Graphic Era (Deemed to be University), Dehradun 248002, India
e-mail: amu.santosh@gmail.com

V. Singh
e-mail: vijaysingh_agra@hotmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_22

is a concept of early 60's and lots of research have been published on this topic. In the current era of Big Data, still, the concepts play a vital role. Kenneth et al. [2] performed a data analysis on school going students using multilevel LCA. They did a drug use and a health-related survey among the school going students and generate the meaningful groups to healthy environments and detect the school level behavior. Latent class growth analysis [3] is used to analyze the physical activities among women. Most of the data sample shows that very less physical activity. Suleyman Cetintas et al. [4] proposed student performance evaluation system using probabilistic latent class model . For demonstrating the methodology, they used real-world, large-scale dataset and considered almost all information attribute regarding students to identify unobserved groups and educational contents. Another interesting study was done by Jeonghoon Ahn [5] for classifying patients using LCA. Their work focuses on antipsychotic drug therapy and they investigate Sociodemographic, resource and Clinical utilization. Analysis of measurement error in consumer expenditure survey was done by Clyde [6]. They use approximately 30,000 cases in the tenure of 4 years (from 2005 to 2009) and analyzed subgroups to measurement error in consumer expenditure report. Nema Dean and Adrian E.Raftery [7] proposed the method, how to select the suitable variable for LCA? They used head-long search algorithm for selecting clustering variable and they demonstrated their work on two real-world datasets. Christopher DuBois and Padhraic Smyth et al. [8] analyzed sequence of dyadic interaction between individuals in the social network scenario. They used two algorithms for this expectation—maximization and Markov chain Monte Carlo based sampling and evaluated the system on large-scale dataset having attribute of receiver, sender, and event type. They investigate real-life issues like missing data. Jaewoo Choi and Woonsun did a runtime detection of Motives for using social media usage among the UG students. They applied LCA and multinomial logistic regression on the dataset collected from 840 Korean UG students. The main finding of them is that, the Motives are a very important predictor of latent class membership of social media usage [9]. George B. Ploubidis and his team-mates used LCA in a different scenario altogether in the analysis of GHQ-28 social dysfunction items. Their work focuses on person-centered analysis model, classes an individual by a dichotomous latent variable, and investigates social functioning classification [10]. Another interesting domain where LCA is applied is by Arthur L.C. Antoine and Keith R. Molenaar, in the field of highway design and construction project categorization [11]. Their analysis explores the characteristics of the project including the facility type, the size of the project, and the complexity of the project. Their analysis shows the effective and defensible implementation of LCA of project classification.

1.1 Review of Latent Class Analysis (LCA)

Here we present the brief review for better understanding of this article. Latent class is a statistical approach and it is a subset of structural equation modeling? The main motto behind the Latent Class analysis is to and out the hidden groups from the observed dataset using the log-likelihood [12]. Latent class Model analogous with the scenario in which P is an Observed dichotomous variable consisting I groups (a = 1, 2, 3, …, A), Variable Q is an observed dichotomous variable consisting J groups (b = 1, 2, 3, …, B), While the variable R is a latent dichotomous or unobserved variable consisting S groups (c = 1, 2, 3, …, C). Let π_{abc}^{PQR} represents the join probability in group a on variable A, similarly in group b on variable Q and in c on variable R. Let $\pi_{ac}^{\bar{P}R}$ representing the conditional probabilities in class a on variable P, $\pi_{bc}^{\bar{Q}R}$ representing the conditional probabilities in group in group b on variable Q, similarly Let π_{c}^{R} representing the conditional probabilities in group c on variable R. The Latent Class Model by McCutcheon [12] can be represent in this condition as follows:

$$\begin{aligned} &\pi_{abc}^{PQR} = \pi_{c}^{R}\pi_{ac}^{\bar{P}R}\pi_{bc}^{\bar{Q}R} \\ &\text{For a} = 1, \ldots, \text{A}; \text{b} = 1, \ldots, \text{B}, \text{c} = 1, \ldots, \text{C} \end{aligned} \tag{1}$$

According to the above-mentioned model P and Q are conditionally independent to each other, given the group level on variable R, shown in Eq. 2.

$$\pi_{abc}^{\bar{P}\bar{Q}R} = \frac{\pi_{abc}^{PQR}}{\pi_{c}^{R}} = \pi_{ac}^{\bar{P}R}\pi_{bc}^{\bar{B}R}, \tag{2}$$

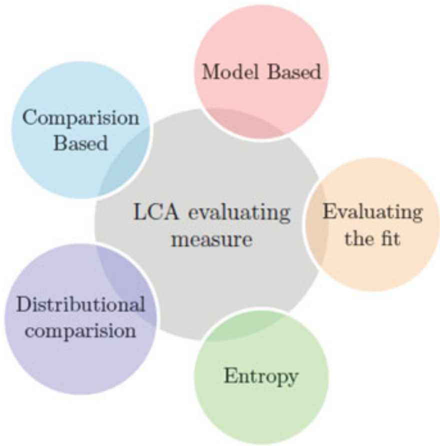
where $\pi_{abc}^{\bar{P}\bar{Q}R} = \frac{\pi_{abc}^{PQR}}{\pi_{c}^{R}}$ is the conditional probability in group a on variable P and in group b on variable Q, given that observation in group c on variable R.

The evaluating measures of LCA are important and that are shown in Fig. 1. The measures are Evaluating the fit of an LCA, Model-based measures of fit, Model comparison measure, Distributional comparisons and entropy. In the model-based measures, log-likelihood and the Chi-square test are involved. In the model comparison measure Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are used. Distributional comparison consists predicted item means and predicted item covariance. Entropy is used to evaluate the classification certainty (Macready and Dayton 1977). X^2 test is using to compare the sets of response patterns with the response patterns expected and the Chi-square test perform better in the scenario where the sample size is larger, and the number of manifest variable are less. If the too many cells consisting zeros, then the Chi-square test is not valid. The log-likelihood is nothing but the model formulation over the joint distribution of the data and used to observe responses of each parameter. The Akaike Information Criterion (AIC) is used to evaluate the goodness of fit of a model that considers the number of model parameters(q). AIC evaluating formula is shown in Eq. 3.

$$AIC = 2q - 2logL \tag{3}$$

There is other information criterion besides mentioned above and most of them are entropy-based.

Fig. 1 Evaluating measure of LCA



2 Materials and Methods

Here, our main concern how to detect or predict the Good performer or Bad performer in the context of Hotel based on attribute categories. Therefore, statistical modeling approach LCA is applied to Trip Advisor dataset and the promising results. The subsequent section is organized as

- Trip Advisor Dataset Description.
- Preprocessing of Data.
- Latent Class Analysis.
- Comparisons of Results.

2.1 Preprocessing of Data

One of the main challenges while doing a data analysis task in a huge data an efficient preprocessing of data is required, to accomplishing this task Text Mining(tm) R package and python program are used JSON format dataset is converted into comma separated file. Rating are given in the numeric form with a range 1–5 percentage Value P_v is calculated by the formula mentioned in Eq. 4.

$$P_v = \frac{\sum_{i=1}^{n} X_i * 100}{n * 5}, \tag{4}$$

where n is the number of ratings available. Sample data is shown in Table 1.

These percentage value p_v six decision class level are embedded with the table. These classes are with their numeric values Excellent (1), very good (2), good (3), moderate (4), below moderate (5), worst (6). Then normalized the dataset to dichoto-

Table 1 Sample data

Hotel name	Service	Cleanliness	…
Best Western	91.5	87.6	…
Grace Inn Phoenix	49.45	68.23	…
Holiday Inn	78.56	69.3	…

Table 2 A sample data with dichotomous values

Hotel name	Service	Cleanliness	…
Best Western	1	1	…
Grace Inn Phoenix	2	2	…
Holiday Inn	1	2	…

Table 3 Conditional probabilities of each class

Conditional probabilities of each class			
Service	Class 1 Class 2	Pr(1):0.9670 Pr(1):0.1397	Pr(1):0.0330 Pr(1):0.8603
Cleanliness	Class 1 Class 2	Pr(1):0.9979 Pr(1):0.3108	Pr(1):0.0021 Pr(1):0.6892
Value	Class 1 Class 2	Pr(1):0.9496 Pr(1):0.0862	Pr(1):0.0504 Pr(1):0.9138
Quality	Class 1 Class 2	Pr(1):0.9408 Pr(1):0.1534	Pr(1):0.0592 Pr(1):0.8462
Room	Class 1 Class 2	Pr(1):0.7147 Pr(1):0.1667	Pr(1):0.2853 Pr(1):0.8339
Location	Class 1 Class 2	Pr(1):0.8922 Pr(1):0.0475	Pr(1):0.1078 Pr(1):0.9525

mous values. If the percentage value P_v is greater than 70 then the value replaced by 1 otherwise replaced by numeric value 2. The sample output is shown in Table 2.

2.2 *Latent Class Analysis (LCA)*

In this section Latent Class Analysis is applied to find the hidden unobserved classes within the Trip Advisor dataset. After getting the normalized data from the preprocessing stage, applied Latent Class Analysis using the R library poLCA. Here the two cases are considered to implementation of latent class analysis. In the first case two-classes latent classes are used and the conditional probabilities are shown in Table 3, while in the second case two-classes latent class regression using Decision attribute as a covariate is used to detect the groups as “Good” versus “Bad”. Conditional probabilities of the second case are shown in Table 4.

Table 4 Conditional probabilities of each class

Conditional probabilities of each class			
Service	Class 1 Class 2	Pr(1):0.9639 Pr(1):0.1346	Pr(1):0.0361 Pr(1):0.8654
Cleanliness	Class 1 Class 2	Pr(1):0.9970 Pr(1):0.3035	Pr(1):0.0030 Pr(1):0.6965
Value	Class 1 Class 2	Pr(1):0.9413 Pr(1):0.0900	Pr(1):0.0602 Pr(1):0.9100
Quality	Class 1 Class 2	Pr(1):0.9398 Pr(1):0.1450	Pr(1):0.0602 Pr(1):0.8550
Room	Class 1 Class 2	Pr(1):0.7181 Pr(1):0.1529	Pr(1):0.2819 Pr(1):0.8471
Location	Class 1 Class 2	Pr(1):0.8870 Pr(1):0.0461	Pr(1):0.1130 Pr(1):0.9539

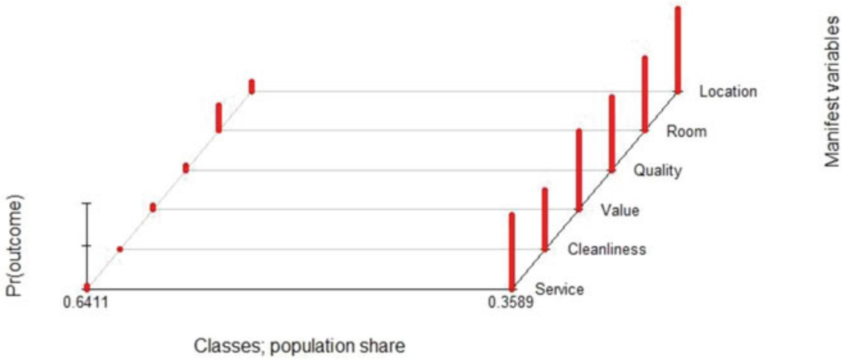


Fig. 2 Classes; population share

Estimated class population shares are 0.6411 and 0.3589. The predictive class membership are 0.6389 and 0.3611 shown in Fig. 2. There are total 9256 observations are taken and the number of estimated parameters are 13. The residual degree of freedom we get is 50, while the log-likelihood is −22678:26. The AIC and BIC value are 45382.53 and 45475.26 respectively. The log-likelihood ratio G^2 is 1930:829 and Chi-square (X^2) value is 2495:23. Decision attribute is considered as a covariate to predict class membership as “Good” versus “Bad”. Estimated class population shares are 0.6457 and 0.3543. The predictive class membership is 0.6428 and 0.3572.

In the second case latent class regression is applied and the total number of observations 9256 is taken. Number of estimated parameters are 14 including Decision attribute. Residual degree of freedom is 49, and the maximum log-likelihood is −19068. The AIC and BIC value are 38165:08 and 38264:94 respectively. The Chi-square (X^2) value is 2467.341. Fit for the two latent classes are shown in Table 5.

Table 5 Fit for two latent classes

	Fit for two latent classes			
	Coefficient	Std. err	t value	Pr(>|t|)
(Intercept)	−10.77433	0.31213	−34.515	0
Decision	2.90822	0.08479	34.300	0

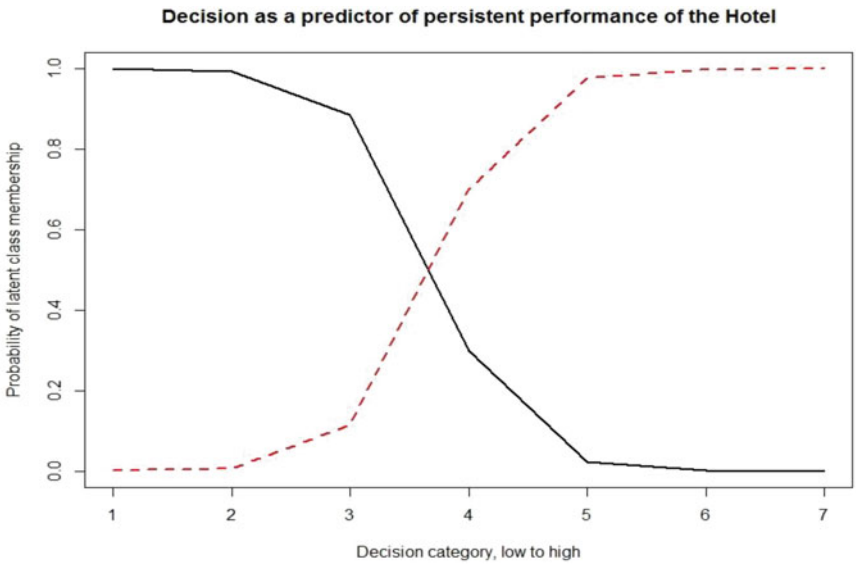


Fig. 3 Class distribution (on training data)

3 Results and Discussion

This study examines the Trip Advisor dataset on the various attribute using the latent class analysis. Results are promising, Fig. 3 shows the perfect inverse graph and the point where the two groups intersect each other (an equilibrium point). A point where there is uncertainty that the object belongs to "Good" or "Bad" class. Figure 3 shows the class distribution (on training data), Fig. 4 shows the class distribution (on testing data), while Fig. 5 shows the comparison on the test cases.

4 Conclusion

Predicting Hotel performance is an important task. This paper proposed a system that is capable enough to predict "Good" or "Bad" Hotel based on attributes using statistics modeling Latent Class Analysis. The systems perform better when there are sufficient data. It is observed that proposed LCA method produced better results on

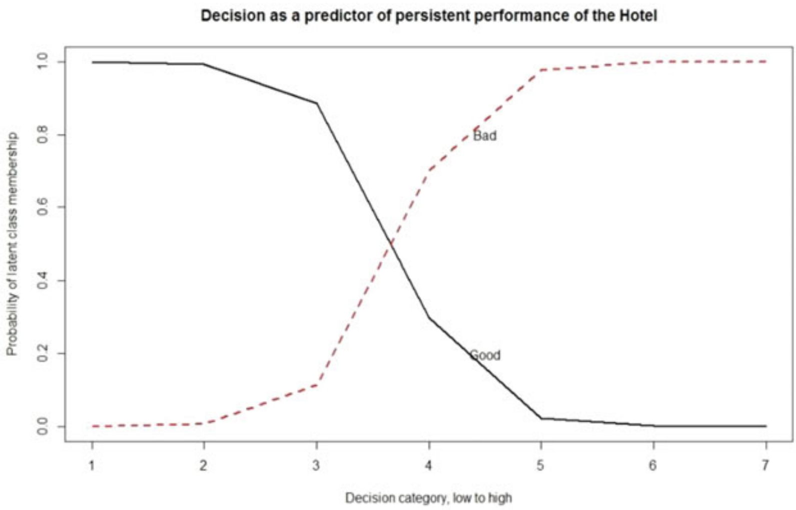


Fig. 4 Class distribution (on testing data)

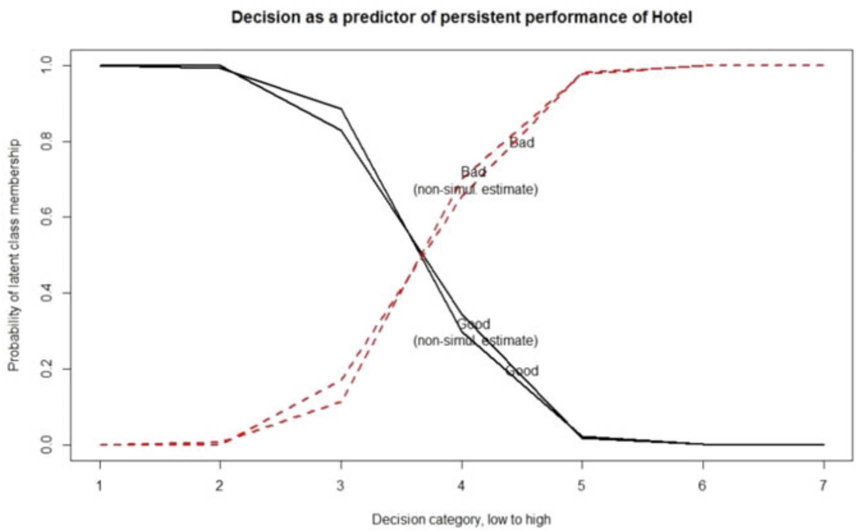


Fig. 5 Comparison on test cases

the function ally independent variables. The proposed methodology can be applied to another similar type of datasets.

References

1. Zhang, N.L.: Hierarchical latent class models for cluster analysis. J. Mach. Learn. Res. **5**(2004), 697–723 (2004)
2. Allison, K.R., Adlaf, E.M., Irving, H.M., Schoueri-Mychasiw, N., Rehm, J.: The search for healthy schools: a multilevel latent class analysis of schools and their students. Prev. Med. Rep. **4** (2016)
3. Kim, Y., Kang, M., Tacon, A.M., Morrow Jr., J.R.: Longitudinal trajectories of physical activity in women using Latent class growth analysis: the WIN study. J. Sports Health Sci. **17** (2016)
4. Cetintas, S., Si, L., Xin, Y.P., Tzur, R.: Probabilistic latent class models for predicting student performance. In: International Conference on Information and Knowledge Management, pp. 1513–1516 (2013)
5. Ahn, J., McCombs, J.S., Jung, C., Tim, J., McDonnell, D., Ascher-Svanum, H., Edgell, E.T., Shi, L.: Classifying patients by antipsychotic adherence patterns using latent class analysis: characteristics of nonadherent groups in the California Medicaid (Medi-Cal) program. In: International Society for Pharmacoeconomic and Outcomes Research (ISPOR), 1098-3015/08/48 (2008)
6. Tucker, C., Meekins, B., Biemer, P.: Latent Class Analysis of Measurement Error in the Consumer Expenditure Survey. Section on Survey Research Methods-JSM2011 (2011)
7. Dean, N., Raftery, A.E.: Latent class analysis variable selection. Ann. Inst. Stat. Math. 11–35 (2010)
8. DuBois, C., Smyth, P.: Modeling relational events via latent classes. In: 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2010)
9. Choi, J., Kang, W.: A dynamic examination of motives for using social media and social media usage among undergraduate students: a latent class analysis. Soc. Behav. Sci. **131**(2014), 202–210 (2014)
10. Ploubidis, G.B., Abbott, R.A., Huppert, F.A., Kuh, D., Wadsworth, M.E.J., Croudace, T.J.: Improvements in social functioning reported by a birth cohort in mid-adult life: a person-centred analysis of GHQ-28 social dysfunction items using latent class analysis. Pers. Individ. Differ. **42**(2007), 305–316 (2007)
11. Antoine, A.L.C., Molenaar, K.R.: Latent class analysis for highway design and construction project categorization. In: International Conference on Sustainable Design, Engineering and Construction, vol. 145, pp. 1314–1321 (2016)
12. McCutcheon, A.C.: Latent Class Analysis. Sage Publications, Beverly Hills (1987)

Analysis of Probabilistic Models for Influence Ranking in Social Networks

Pranav Nerurkar, Aruna Pavate, Mansi Shah and Samuel Jacob

Abstract Influence is a phenomenon occurring in every social network. Network science literature on Influence ranking focuses on investigation and design of computational models for ranking of nodes by their influence and mapping the spread of their influence in the network. In addition to this contemporary literature seeks efficient and scalable influence ranking techniques that could be suitable for application on massive social networks. For this purpose joint and conditional probabilistic models could be a way forward as these models can be trained on data rapidly making them ideal for deployment on massive social networks. However identification of suitable predictors that may have a correlation with influence plays a major role in deciding the successful outcome for these models. The present investigation proceeds with the intuition that interaction is positively correlated with influence. Furthermore, through extensive experimentation it identifies a joint probabilistic model and trains it on interaction characteristics on nodes of a social network for influence ranking. A qualitative analysis of these models is presented to highlight its suitability.

Keywords Social network analysis · Social influence analysis
Network centrality · Influence Ranking

P. Nerurkar (✉)
Department of CE & IT, VJTI, Mumbai, India
e-mail: pranavn91@gmail.com

A. Pavate
Department of CE & IT, Atharva CoE, Mumbai, India
e-mail: gavkare@gmail.com

M. Shah
Department of CE & IT, Rizvi CoE, Mumbai, India
e-mail: mansishah928@gmail.com

S. Jacob
Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu, Rajasthan, India
e-mail: samueljacob@atharvacoe.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_23

1 Introduction

Social influence is referred to as the behavioural change or alternation in performance of actions of an individual brought about due to interchanges conducted with other individuals. It has known to be the causal link for other well documented phenomena seen in social networks such as social competition, peer pressure, homophily, information spread, network evolution. Influence is also important in constraining the flow of dynamics within a network [1].

In online social networks, social influence may depend on factors such as strength of relationships between nodes, distance between the nodes, number of paths for traversal from a node to its neighbors and characteristics of the individuals in the network [2]. However, for the purpose of developing a computational model that shall measure influence quantitatively and qualitatively, a statistic based measure is required. Statistic based measures proposed in the literature for measuring influence are centrality, between-ness, closeness, decay etc. These methods focused only on the structural characteristics of a social network to calculate the influence of nodes in it. Influence Maximization technique goes beyond such simple statistical measures and provides an alternate method to measure influence. However it is a subset selection problem and hence is fundamentally different from Influence ranking which is a measurement problem. Hence it is not the focus of the present inquiry.

Joint and conditional probabilistic frameworks could be the suitable methods to measure influence. This is because such techniques have been experimentally verified on wide range of scenarios. The advantage of these methods is that they can be trained rapidly on data. This makes them suitable for deployment on massive social networks such as Twitter, Facebook etc. However, a key aspect that determines the success of these methods is the selection of suitable predictors for training them. This inquiry selects statistics about user interactions for training four probabilistic models. Extensive experiments are then performed using these frameworks on Twitter data. Standard evaluation metrics are used to select the most optimal model out of these. The aim of this work is to extend the literature on use of probabilistic models to rank nodes on influence.

2 Review of Literature

The edge and node measures for calculating centrality are techniques that rely on structural features which ignoring attribute level data of a node or interaction characteristics of a node in the network. To overcome these drawbacks Influence Maximization technique and Probabilistic Generative models were proposed.

2.1 Quantifying Influence

A formal computational definition for influence is proposed in [3]. The nodes are x_i and x_j and t and $t-1$ are the time instants, a_{ij}^t denotes value of the adjacency matrix at time instant t for nodes i and j.

$$\frac{p((x_i^t, x_j^t) > (x_i^{t-1}, x_j^{t-1})|a_{ij}^{t-1} = 0, a_{ij}^t = 1)}{p((x_i^t, x_j^t) > (x_i^{t-1}, x_j^{t-1})|a_{ij}^{t-1} = 0)} \tag{1}$$

The numerator is the conditional probability that two nodes that were not linked at instant $t-1$ are linked at t have seen increase in their similarity. The denominator calculates the conditional probability that two nodes that were not linked at $t-1$ see an increase in their similarity at t as compared to $t-1$. This method however does not differentiate the influence from various angles [2].

2.2 Influence Maximization (IM)

The formal definition of the problem is: Given a social influence graph $G(V, E)$ with V, E representing vertices (individuals) and edges (social relationships) respectively. $P(u, v)$ is the probability that v is activated by an already active node u sharing a directed edge (u, v). Independent cascade model allows a small sets of seed nodes to be activated. Then a node u can activate its neighbor v with probability $p(u, v)$. Influence spread is denoted as the maximum number of nodes activated. The problem is to maximize the influence spread [4]. IM is an NP-hard problem and hence a greedy approximation algorithm is used that can theoretically guarantee influence spread is within 63% of the optimal influence spread. However, the greedy algorithm requires the evaluation of the influence spread given a seed set. This step is time consuming.

In contrast to this line of work, statistics of interaction by nodes were also explored for measuring their influence in the network. Interaction occurs in the network through activity performed by the node, response received from its neighbors on such an activity, propagation of activity further throughout the network, activating reaction from nodes not directly connected to each other etc. Klout rank is one such measure that ranks nodes in a network as per their influence based on their statistics of interaction [5]. These statistics are collected from multiple social networks of which the node is a participant. Klout rank uses a feature vector of 3600 attributes which are statistics of user interaction on the social platform and analyzes interactions between the user and other participants on the social network to generate a score. Alexy et al. have investigated use of machine learning for calculating influence ranks based on data generated from social network activity by the users [6]. Behnam et al. modelled influence of a node in a network based on metrics such as number of followers and ratio of affection [6]. The purpose of the current inquiry is to advance the existing knowledge in the field of influence analysis by providing experimental evaluation of conditional and joint probabilistic learning techniques to model influence of nodes on social networks.

3 Mathematical Model

Learning techniques build a function $g \in G$, where G represents the hypothesis space. Function $g : X \rightarrow Y$ maps the input space X to the output space Y. If F denotes the space of scoring functions f such that $f : X * Y \rightarrow R$ then g is as shown in Eq. 2.

$$g(x) = \arg\max_{y} f(x, y) \tag{2}$$

Then a suitable conditional probability model or joint probability model is chosen for mapping f to g. For a conditional probability model, g is calculated as $P(y|x)$ and for a joint probability model g is $P(x, y)$. A method to choose the appropriate model is structural risk minimization in which a regularization penalty called L_2 norm which is $\sum_j \beta_j^2$ is incorporated in to the cost function $J(\theta)$ to minimize over-fitting.

3.1 Learning Technique for Ranking Influential Nodes

Conditional probability model or joint probability models are then used in the below technique to identify the most suitable model amongst them for comparing influences and ranking nodes as per their influence.

Algorithm 1: Influence Ranking model

```
Set initial seed for random numbers;
Set the training control values;
Set the tuning grid for parameter search;
for each parameter set do
    for each re-sampling iteration set do
        hold out specific samples;
        Pre process the data (Center and Scale);
        Fit the model on the remaining samples;
        Predict the held out samples;
    Calculate the average performance across held out predictions;
Determine the optimal parameter set;
Fit the final model to all the training data using optimal parameter set ;
```

Table 1 Description of the data-set

Sr. No.	Name	Description
1	Training set size	5500
2	Test set size	5952
3	Feature vector	8
4	Classification	Binary

Table 2 Attributes in the feature vector

Sr. No	Feature list	Sr. No	Feature list
1	Follower count	5	Following count
2	Listed count	6	Mentions received
3	Retweets received	7	Mentions sent
4	Retweets sent	8	Posts

4 Experimental Study

4.1 Data-Set

Data-set used for the experiment consists of features extracted about user interaction characteristics (Tables 1 and 2) from Twitter—an online social network [6]. Twitter is a graph $G = (V, E)$ with n nodes having k attributes which have been formed using its interaction characteristics. These have been collected from the nodes activity observed over an online social network. Any two nodes, A, B of G are picked and a feature vector x_i is built by combining their individual interaction characteristics i.e. $x_i = a_i, b_i$. This is used to build a dataset X that contains training samples of the form $(x_1, y_1)...(x_n, y_n)$ such that x_i is the feature vector of the ith example. The corresponding class label of x_i is y_i which represents the human judgment about which one of the two individuals in x_i is more influential. Thus $y_i \in 0, 1$ such that $y_i = 0$ means first user is more influential and $y_i = 1$ means that the second user is more influential.

Learning models described in Sect. 4.2 are used in the Influence learning strategy proposed in Sect. 3.1. Performance metrics used for selection of the appropriate model are Accuracy, Kappa, Area under the ROC curve, specificity, sensitivity, Log-loss, Precision and Recall.

Table 3 Parameters of the Optimal Ex-LM model

Model	nhid	actfun
Ex-LM-1	14	Radial basis

Table 4 Parameters of the Optimal Rf model

Model	mTry	Metric
Rf-1	2	Accuracy
Rf-2	1	ROC
Rf-3	3	Logloss
Rf-4	7	AUC

4.2 Result

4.3 Artificial Neural Network: [7, 8]

This model is a single hidden layer feed-forward neural network whose tunable parameters are number of hidden units ($nhid$) and activation function ($actfun$) chosen using the random hyper-parameter optimization [9]. The optimal model was chosen by fivefold cross validation repeated 5 times on the basis of Accuracy metric (other metrics aren't calculated for this model as it doesn't give class probabilities) (Table 3).

4.4 Random Forests (Rf): [10]

Random forests were the next model trained on the interaction characteristics. Tunable parameter is the number of trees ($mTry$) for this model. Four different optimal models were selected one each for the metrics Accuracy, Area under ROC curve (ROC), Area under Precision-recall curve (AUC) and logloss. fivefold cross validation was used with 5 times repeat (Table 4).

4.5 Stochastic Gradient Boosted Trees (GBM): [11]

Stochastic Gradient Boosted Trees were the next model trained on the interaction characteristics. Tunable parameter is the boosting iterations ($n.trees$), maximum tree depth ($interaction.depth$), ($shrinkage$) and minimum terminal node size ($n.minobsinnode$) for this model. Four different optimal models were selected one each for the metrics Accuracy, Area under ROC curve (ROC), Area under Precision-recall curve (AUC) and logloss. fivefold cross validation was used with 5 times repeat (Table 5).

Table 5 Parameters of the Optimal GBM model

Model	n.trees	interaction.depth	shrinkage	n.minobsinnode	Metric
GBM-1	2695	2	0.06	12	Accuracy
GBM-2	190	1	0.42	13	ROC
GBM-3	32	2	0.56	12	Logloss
GBM-4	196	10	0.32	20	AUC

Table 6 Parameters of the Optimal xgbTree model

Model	nrounds	max-depth	eta	gamma	col-bt	min-cw	subsample	Metric
xgbTree1	573	10	0.09	5.01	0.64	19	0.91	Accuracy
xgbTree2	816	1	0.4	3.02	0.55	11	0.49	ROC
xgbTree3	901	1	0.06	7.65	0.64	8	0.72	logloss
xgbTree4	175	2	0.05	2.7	0.59	13	0.81	AUC

4.6 Extreme Gradient Boosted Trees (xgbTree): [11]

Extreme Gradient Boosted Trees model was trained on the interaction characteristics. Tunable parameter are the Boosting Iterations ($nrounds$), Max Tree Depth ($max - depth$), Shrinkage (eta), Minimum Loss Reduction ($gamma$), Sub-sample Ratio of Columns ($col - bt$), Minimum Sum of Instance Weight ($min - cw$), and Sub-sample Percentage ($subsample$) for this model. Four different optimal models were selected one each for the metrics Accuracy, Area under ROC curve (ROC), Area under Precision-recall curve (AUC) and log-loss. fivefold cross validation was used with 5 times repeat (Table 6).

Optimal models selected from the above procedure were selected evaluated on the test data. Based on the evaluation metrics it can be inferred the xgbTree models fit the test data better than other ML techniques applied on the training dataset. xgbTree models are ensembled predictors of balanced decision trees and also have the advantage of not overfitting the data.

Table 7 demonstrates that Extreme Gradient Boosted Trees (xgbTree) were found to be the most appropriate probabilistic models amongst those reviewed in this study for influence ranking on the Twitter dataset. xgbTree model supports parallelization of tree construction, cache optimization of data structures and algorithm to make best use of hardware, distributed computing and out-of-core computing for very large datasets that do not fit into memory for training very large models using a cluster of machines [12].

Table 7 Results of Optimal ML based model on test data

Model	Accuracy	Kappa	ROC	Sensitivity	Specificity
Ex-LM-1	0.66	0.32	–	–	–
Rf-1	0.77	0.54	0.77	0.77	0.78
Rf-2	0.77	0.54	0.77	0.76	0.77
Rf-3	0.77	0.54	0.77	0.76	0.78
Rf-4	0.77	0.54	0.77	0.76	0.78
GBM-1	0.78	**0.57**	0.78	0.76	0.79
GBM-2	0.78	0.55	0.78	0.77	0.78
GBM-3	0.78	0.56	0.78	0.76	0.79
GBM-4	0.76	0.52	0.76	0.74	0.77
xgbTree-1	**0.79**	**0.57**	0.79	**0.77**	**0.8**
xgbTree-2	0.77	0.53	0.77	0.75	0.78
xgbTree-3	0.78	**0.57**	0.78	**0.77**	0.79
xgbTree-4	**0.79**	**0.57**	**0.79**	**0.77**	**0.8**

5 Conclusion

This inquiry was conducted on the intuition that interaction occurs between entities of any social network and this could play an important role in modelling of influence. Hence statistics related to interactions made by entities in a social network were utilized to model influence and rank the entities as per their influence. Joint and conditional probability based techniques were used such as Neural networks, decision trees, ensembled predictors to conduct the inquiry. Based on the evaluation of the performance of the learning techniques on the data it was concluded that Extreme gradient boosted trees were the most suitable amongst the techniques used. The key advantage of this technique over other methods would be that it could scale well over massive online networks compared to other influence ranking models.

References

1. Aggarwal, C.C.: An introduction to social network data analytics. In: Social Network Data Analytics, pp. 1–15 (2011)
2. Sun, J., Tang, J.: A survey of models and algorithms for social influence analysis. In: Social Network Data Analytics, pp. 177–214 (2011)
3. Scripps, J., Tan, P.-N., Esfahanian, A.-H.: Measuring the effects of preprocessing decisions and network forces in dynamic network analysis. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 747–756. ACM (2009)
4. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146. ACM (2003)

5. Rao, A., Spasojevic, N., Li, Z., DSouza, T.: Klout score: measuring influence across multiple social networks. In: 2015 IEEE International Conference on Big Data (Big Data), pp. 2282–2289. IEEE (2015)
6. Nerurkar, P., Bhirud, S.: Modeling influence on a social network using interaction characteristics. Int. J. Comput. Math. Sci. 152–160 (2017)
7. Zurada, J.M.: Introduction to Artificial Neural Systems, vol. 8. West St. Paul (1992)
8. Huang, G.-B., Zhu, Q.-Y., Siew, C.-K.: Extreme learning machine: a new learning scheme of feedforward neural networks. In: 2004 IEEE International Joint Conference on Neural Networks, 2004. Proceedings, vol. 2, pp. 985–990. IEEE (2004)
9. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res. **13**, 281–305 (2012)
10. Breiman, L.: Random forests. Mach. Learn. 5–32 (2001)
11. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232 (2001)
12. A gentle introduction to XGBoost for applied machine learning, Sep 2016

Smart City Project Management System Using Cloud

Revati M. Wahul and Santosh S. Lomte

Abstract Massive increase in population around the world and the advent of more and more number of people moving to cities for livelihood has increased the demand for better transportation and infrastructure. It has given rise to conflicts between multiple smart city services and demands the better project management. Here, we are putting forward a fresh approach for smart city project management using the live data feed through which we detect the conflict in real time, which in turn helps the authorities for better decision-making.

Keywords Smart city services · Cloud computing · Data mining
Project management

1 Introduction

Everlasting growth of road framework, vehicles, and their movement problems have transformed into a boundless issue in nearly all urban regions around the globe. In the midst of crest hours, it is mind-bogglingly typical that an emergency vehicle can jam in long vehicle line for a long duration. This negatively impacts the general well-being and national economies [1]. Due to this, it is possible the rescue operation may be badly affected in the emergency hours. A decent administration method in post fiasco circumstance necessitates informing each and every concerning vehicle and also on the route of the emergency vehicle so they can clear out and offer a way to the emergency vehicles.

R. M. Wahul (✉)
Computer Science and Engineering, Dr. Babasaheb Ambedkar
Marathwada University, Aurangabad, India
e-mail: reva.wahul@gmail.com

S. S. Lomte (✉)
VDF School of Engineering & Technology, New MIDC, Airport Road, Latur, India
e-mail: drsslomte@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_24

Lately, the Internet of Things (IoT) is fixated on structures, conventions, frameworks, and administration for the sensible interconnection of diverse things, to establish sending, and a development of noteworthy worth included administrations [2]. The weight of the IoT things, administrations, and stages is sustained by distributed computing and cloud computing stages. With the IoT being a multidisciplinary ecosystem, it is right now being utilized in regards to circumstances asking for consistent data getting ready and providing responses, for example, related and self-decision vehicle circumstances [3].

In this paper, we simplify the fundamentals of what outlines a smart city, which we require as a city in ease of use and in which the ICT is mixed with standard frameworks, and supervised and synchronized using new digital advancements. With these advantages, citizens, government authorities and other organizations can work together to understand the city in a better manner. We put our idea for building up new information of urban issues, for instance, response for emergency services.

2 Literature Survey

Past centuries witnessed the monstrous growth of industrialization and as an effect of its humongous urbanization, which in turn makes ways for new and incredible cities—at first in Europe, and then worldwide. The new employment possibilities on horizon have resulted in the massive immigration of rural population from rural villages into these new but unplanned cities. In 1900, just 13% of world populations were in the city area; and by 1950, the mark reached 29% [1].

Cities are expanding swiftly and with the expanding urban populaces, the development should be accelerated all the more effectively. Developing cities have restrictions on land usages, and they are defining their new lifestyles and traditions, which are beyond the authority reach to get easily controlled. City boundaries are tough to characterize being based on economics more specifically. It often causes animosity among various authorities involved [3].

The smart city vision must be endorsed through open, controlled, and consolidated urban working frameworks to maintain a strategic distance from traders forcing business models and offer an unchecked data to all citizens instead of removing virtual gated assemblies and corporate enclaves. Urban execution directly depends not only on the city's advancement of smart capital but also on the openness and nature of human and social capital.

The critical test that we and the rest of the society go up against is gripping as we develop new digital innovations; we use those same advancements to think about the systems, their application, usage, and impact on society.

3 Problem Statement

The proposed work aims to develop a smart city management system wherein real-time multiple service operation and the conflict detection will be available.

4 System Overview

Smart city platform has diverse troubles that can be of distributed nature in terms of architecture and they support us to see distinctive service challenges to settle them in a collaborative way. Other computing architecture can similarly be improved on a very basic level to grow the confounding and heterogeneous architecture. Following this, we illuminated an instance of one of the architectures by investigating the resource model. We in like manner exhibit various sorts of use cases, which can be associated with smart city architecture and to the known computer architecture.

A. ***Resource Model***

The physical computing provision accessible to smart city project operations incorporates computation and communication sources. Computation sources incorporate equipment apparatuses required to perform computation tasks, while communication implies to the instruments vital for cooperation among jobs executing on other computation nodes. This includes communication bandwidth, latency, network topology, and conceivable safety efforts [4].

B. ***Present Computation Patterns and Use Types***

Diverse sorts of applications can be accommodated by smart city platforms. These applications are extended on resources of the same computing group and furthermore resources of different computing groups depend on the application [4].

Nowadays, increasing the number of vehicles on roads increases the problem of traffic congestion. Due to which it takes more time for an emergency vehicle to reach its destination. To overcome this, relied on all statistics, traffic congestion should be minimized or controlled. So, the proposed system is built in real time. This application works as the lifesaver. Cloud plays an important role in emergency vehicle and all other transportation vehicle or commercial vehicle [4].

5 Implementation

5.1 Big Data Acquisition Unit

For our experiment, an API data feed is used for live sensing of data from remote areas and at runtime, the framework is designed to filter and validate the data and store it to the NoSQL architecture for further processing.

Algorithm I. Filtration and Load balancing algorithm

Input: Live data feed process dataset
Output: filtered data in fixed size block and send each block to processing mechanism
Steps:

1. Filter related data, i.e., processed data. All other unnecessary data will be discarded.
2. Divide the data into appropriate key–value pair.
3. Transmit unprocessed data directly to aggregation step without processing.
4. Assign and transmit each distinct data block of processed data to various processing steps in data processing unit.

Description: This algorithm takes live data and then filters and divides them into segments and performs load balancing algorithm.
In Step 1, related details are filtered out.
In Step 2, filtered data are associated with different key value pairs and each pair is different numbers of sample, which results in forming a data block. In next steps, these blocks are forwarded to be processed by the data processing unit.

5.2 Data Processing Unit

Discarding the unwanted data and preserving the application critical data should be the basic function of any data primary processing architecture and since we are using the cloud to facilitate our processing environment, it is feasible to process and filter massive amount of data in real time.

Algorithm II. Processing and calculation algorithm

Input: Filtered data
Output: Normalized disruption data.
Steps:

1. For each network rail operator performance, categorical data like G for good, A for average is extracted.
2. Normalize the disruption data for all the three modes
3. Persist the data into data store and forward it.

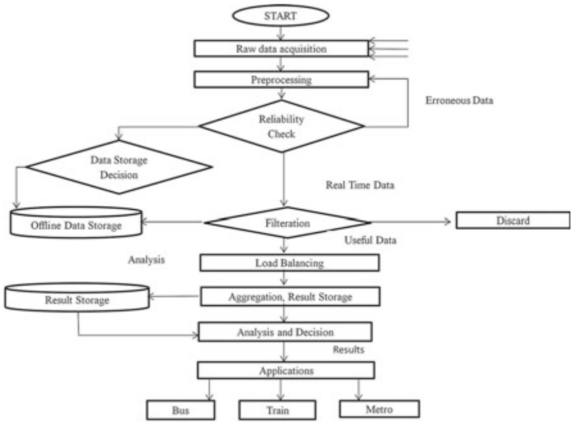


Fig. 1 Smart city data flow diagram

Description: The processing algorithm calculates the results for different parameters against each incoming filtered data and sends them to the next level.
In Step 1, the calculation of good and average along with the trend and furthermore, in the next step, the results are transmitted to the aggregation mechanism.

5.3 Analysis and Decision Unit

This unit is used for further processing and gives the aggregated result of the overall data processed. Various types of application-specific algorithms and business logic classes are written in this unit. The algorithm can change from requirement to requirement and dependent upon the analysis and processing needs of the application, as in our case we are proposing the conflicts on different services in terms of disruption percentage.

Algorithm III. Multimodal summarization algorithm
Input: Normalized disruption data.
Output: Final result summary

1. Gather the data from data store in normalized format.
2. Apply summarization for individual modal pie from the total disruption data capture.
3. Persist the final disruption summary into data store.

Description: Here, the data is collected and the results from each modal are processed against all and then combine, organize, and stores these results in NoSQL database (Fig. 1).

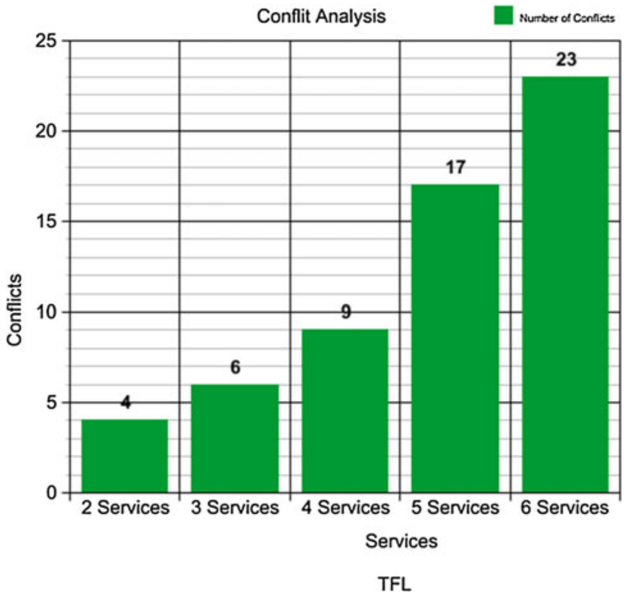


Fig. 2 Service conflicts at peak hour

6 Experimental Results

Experiments were carried on live data stream available on www.tfl.gov.uk data source. There are multiple services data stream available out of which in this system we are taking three different data as inputs (Bus, Train, and Metro). We are using the live feed for data aggregation, and the unwanted data is filtered out and the data is preprocessed for classification step. In the next step, as discussed above, we are tracking the transport services for possible number of conflicts arriving from daily runtime activity, and cloud server Linode has been used as our processing and hosting platform. Our server consists of 4 GB RAM, 2 core processor, and 80 GB HDD with a static bandwidth of 100 Mbps (Fig. 2).

Our system has the following advantages:

1. It improves the efficiency of traveling.
2. Reduces the overhead of the passenger.
3. Reduces disruption.

7 Conclusion

Conflicting services pose genuine security dangers and operational disappointment in a smart city environment. These undertaking centers details on the issue of contentions. In particular, it (i) traces a few qualities of services that contribute toward clashes (conflict), (ii) proposes conflict taxonomy classification as far as starting point of contention.

Furthermore, watchdog architecture is intended for blocking activities from all administrations, and recognizing and settling conflicts.

1. Intelligent transport system is proposed for controlling traffic congestion in favor of emergency vehicle at the time of emergency cases.
2. It provides advance warning in emergency situation and increases the customer satisfaction and also saves critical minutes of a person's life.

In future, almost all cars may have integrated intelligent transport system, which will be in a network, communicating with each other by sending important data about the traffic, arrival of emergency vehicles' alert, etc., for example, live data recording and mapping of emergency vehicles.

References

1. Rahman, H., et al.: Road traffic forecasting through simulation and live GPS-Feed from inter-vehicle networks. In: Proceedings of the GHTC 2012 IEEE, Washington, pp. 36–40
2. Rathore, M., et al.: Real-time big data analytical architecture for remote sensing application. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. (to be published)
3. Corcoran, P., et al.: Mobile-edge computing and the internet of things for consumers: extending cloud computing and services to the edge of the network. IEEE Consum. Electron. Mag. (2016)
4. Pradhan, S., Dubey, A., Neema, S., Gokhale, A.: Towards a generic computation model for smart city platforms. In: 1st International Workshop on SCOPE—GCTC, IEEE, Vienna, Apr 2016

Performance Scaling of Wireless Sensor Network by Using Enhanced OMRA Routing Algorithm

Tanaji Dhaigude, Latha Parthiban and Avinash Kokare

Abstract In modern scenario it has become inherent to employ of Wireless Sensor Networks (WSN) for government and other sectors including defense. Sensor networks can be employ in society, industries, military areas, roads, forests, etc. In a traditional networks it becomes complicated to employ denser node deployment and other problems, e.g., node failure, energy consumption and asymmetric are also prominent. Sensor nodes usually works on battery-powered source and these nodes do not operate for longer time without any manual intervention and it is a very tedious and time-consuming task in forest and defense areas. Hence, it becomes a necessity to reduce energy consumption of sensor, it will also increase energy lifetime. Traditional algorithms like Radio Aware (RA), Distance Source Routing (DSR) and Directed Diffusion (DD) do not solve problems like network connectivity and asymmetric links. To overcome this problem Optimized Mobile Radio Aware (OMRA) technique is demonstrated in this paper.

Keywords Wireless sensor networks · Directed diffusion (DD) · OMRA

1 Introduction

Wireless sensor networks have wide applications in society, industries, military areas, roads, forests, etc. A WSN network is nothing but huge number of small sensor nodes (e.g., heat/humidity sensor) which has different capabilities like sensing capacity, data processing, and communication. These capabilities are necessary for some applications like military applications, monitoring and fire detection. One of the problems

T. Dhaigude (✉)
Bharath Institute of Higher Education and Research, Chennai, India
e-mail: tanajidhaigude@gmail.com

L. Parthiban
Department of Computer Science, Pondicherry University CC, Pondicherry, India

A. Kokare
Faculty, VPKBIET, Baramati, India

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_25

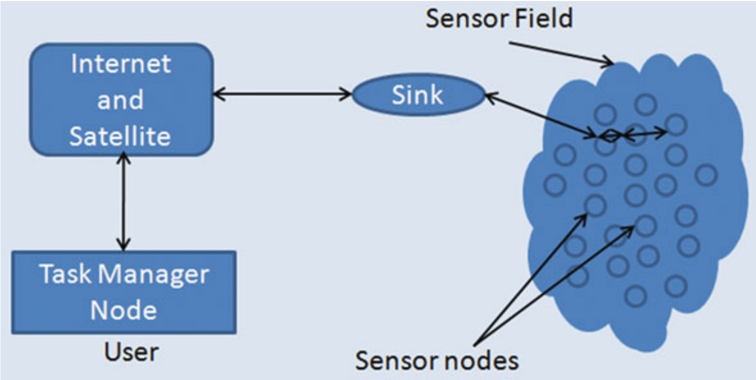


Fig. 1 Typical wireless sensor network

is to replace/recharge sensor node batteries [1]. Because of this minimizing energy consumption is very important.

Motivation for OMRA algorithm is energy consumption in WSN, which is one of the existing problems in WSN. Node failure, asymmetric communication and energy conservation are existing problems in WSN. Different routing algorithms can be used to solve some existing problems of WSN. In proposed OMRA algorithm, network disconnection may occur after removing failed node [2, 3]. Due to limited transmission range of WSN nodes, there will not be direct communication between nodes even after detecting common event [4, 5]. In this paper, enhanced optimized mobile radio aware (OMRA) routing algorithm is proposed. RA routing problem has been handled by OMRA algorithm. By using this algorithm, nodes can be moved to new location which in turn helps in energy saving and also, keeps nodes within communication range (Fig. 1).

In this algorithm, intermediate nodes are choose by considering multiple factors such as node failure, sink node to current node distance, communication radius and airtime cost for each available path. Proposed OMRA algorithm always choose best route amongst source and sink. A nonlinear programming model is used to enhanced the flow rate to sink node and reduce the consumption of energy [6].

2 Related Work

Direct Diffusion (DD) is used in distributed network as it is data centric protocol. To overcome the disadvantages of DD protocol, two totally different routing ways are developed. Main technique focuses on increasing the network period of time by utilizing sensing element node energy properly. In [7–9], the authors planned the ideas of clustering to cut back the energy consumption and increasing the network's period of time. In [10], the minimum remaining energy and gradient depth were used to choose successive hop. In [11], the authors realized that time-sensitive information traffic and load-balanced routing theme was not supported by DD. This motivated them to design a Real Time (RT) filter that provided reduced end to end delay and a Best Effort filter (BE) that gives higher energy potency.

Previous work attempted to solves the problems of Directed Diffusion (DD) protocol and increase the network lifetime. They addressed different kinds of issues, e.g., energy conservation and load balancing, etc. But, in case of lossy WSNs reliable data transmission was not guaranteed in previous work as they considered network as a stable network. Also, previous work did not thought of asymmetric communication link problems in WSN.

The RA routing algorithmic rule was proposed to deal with the issues of lossy WSNs [12] by providing QoS, responsiveness and reduction of overheads in unstable WSNs. It also provides a reliable information transmission for even a lossy WSNs. Steps of RA algorithm:

(1) Get radio information
(2) Select the node
(3) Fast reroute scheme.

RA algorithm has one disadvantage that is network disconnection occurs while removing failing node.

In [13], the optimum device displacement drawback in WSNs that deals with the mobile sensors nodes is introduced. Then on linear programming model proposed in this paper uses totally different situation that permits to adaptively reconfigure the network and repair itself underneath the unpredictable runtime circumstances. In [14], a technique was developed to schedule flows at intervals that maximize the network lifespan. In [6], two general flows- based square measure involving information extraction for relentless energy and less total energy usage was proposed.

2.1 Algorithm Design

Figure 2 shows the flowchart explaining the working of OMRA algorithm. The various steps in the algorithm is given below

- Define a system:
- Let O be OMRA protocol, O = { }
- Identifying the input:

  Let, O1 = {N, Pkr, Pks, t, Rt, Oi}
  O = {O1, …..}
  N = {Nj|j = 1, 2, …} where N is nodes
  Pkr = Packets received
  Pks = Packet sent
  Rt = Data transfer rate from node to sink O1
  Nj = {Imn, Pmn, Fm, Ri, LE, Rc},

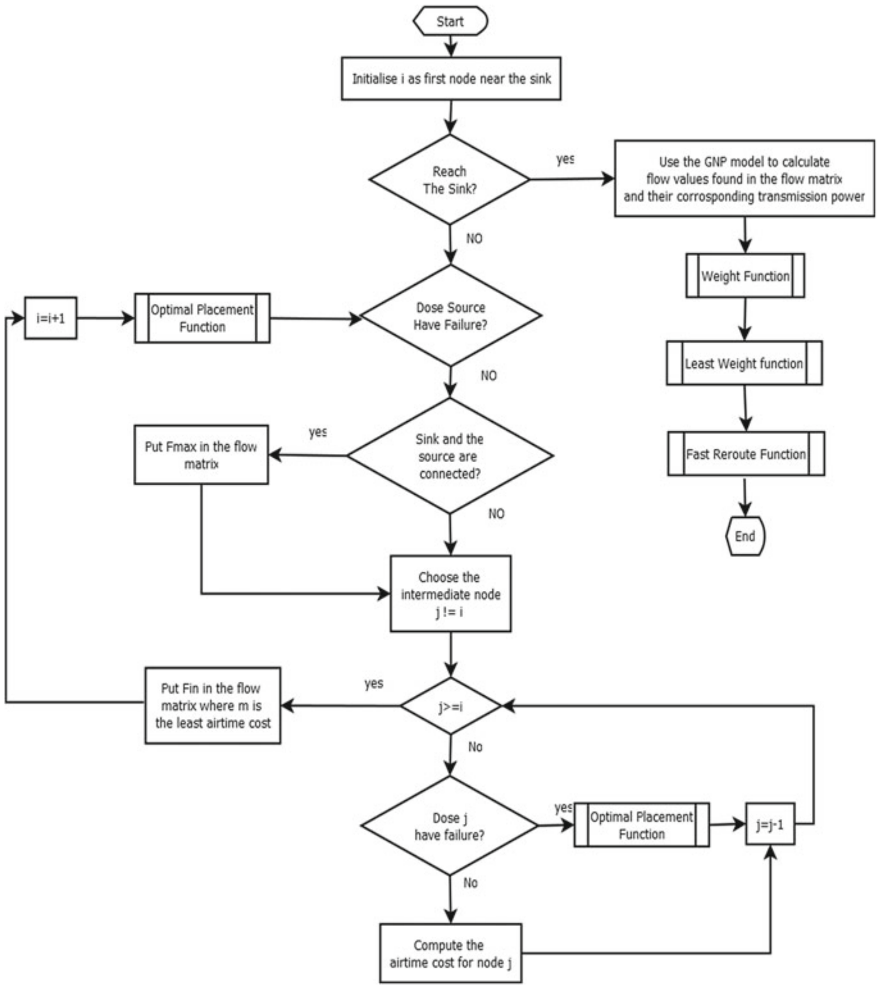


Fig. 2 Proposed algorithm: OMRA

where,

Imn = Node m to node n information flow rate
Pmn = Link m to n transmission power
Fm = Node m's faireness index
Ri = Max source rate for given node
LE = All node's limited energy
Rc = Max communication radius

- Identify the Weights:

$$\text{Wmn} = \text{Fmn} + \text{Pmn} \tag{1}$$

Pmn is Tx power
Fmn is normalized min valu of flow rate

$$\text{Fmn} = \text{fmax} - \text{Imn} \tag{2}$$

where, fmax = Max information extracted from WSN.
In terms of functions,

(1) *Weight Function*:

$$\text{Wmn} = \text{Fmn} + \text{Pmn} \tag{3}$$

Pmn in Tx power

(2) *Least weight function*:

Node_PH = min(Wmn), path from source to sink

(3) *Fast re-route function*:

ifNode_PH = TRUE then select Node_PH else find next least weight Node_PH
The GNP model is used in OMRA algorithm to calculate WSNs flow rate. Advantage of GNP model is that it can balance out the min every consumption and max flow rates. The GNP model hybridizes two nonlinear optimization models. The energy required for each sensor node is addition of transmission energy and reception energy. Equation 4 shows the reception energy and Eq. 5 takes care that the outflow rate is at least as large as in-flow rate for sensor nodes except the aggregators. For aggregator, Eq. 6 is used and σ denotes the data reduction ratio.

$$\max \sum_{k=1}^{n} - \sum_{k=1}^{n+1} P_{ik} + C((1 - \sigma x)\Big(\sum_{k=1}^{n} F_{ik}\Big) + q \sum_{k=1}^{n} f_{ik} \tag{4}$$

Such that

$$\sum_{k=1}^{n} f_{ik} < \sum_{k=1}^{n+1} f_{ik} \tag{5}$$

$$\Big(\sum_{k=1}^{n} f_{ik}\Big)(1 - \sigma x) < \sum_{k=1}^{n+1} f_{ik} \tag{6}$$

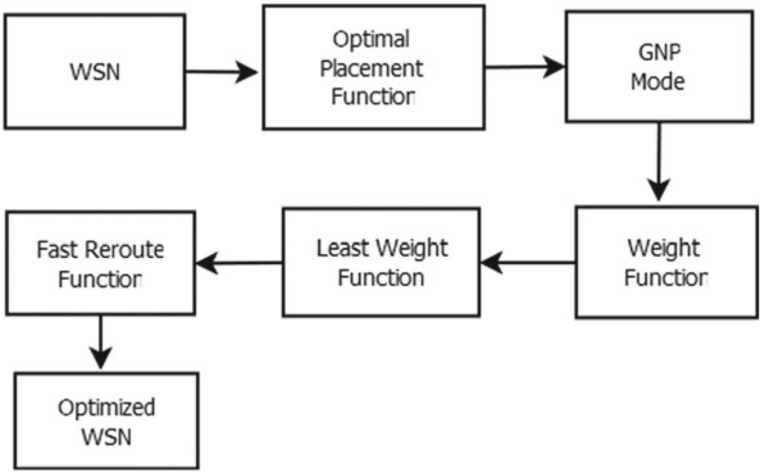


Fig. 3 Data flow architecture

2.2 *Dynamic Programming and Serialization*

The enhanced OMRA algorithm makes use of the following functions to improve the performance of WSN and its data flow architecture is shown in Fig. 3.

1. *Fast Reroute Function*: Fast reroute function algorithm checks the availability of the selected path. Algorithm for fast reroute function is as follows.

   Data: Data Tx-RX path

1. If sensor node N has a data
2. Then N sends RTS signal to node which is intermediate and has highest priority
3. Else wait for incoming data
4. End If loop
5. If N reaches to sink node then return
6. Else if N receives CTS signal before timeout occurs then
7. N sends out the data to node which is intermediate and becomes N
8. Else follows the priority and send it a signal RTS
9. End all if loops.

2. *WEIGHT FUNCTION*: Weight function algorithm is given below is used for finding the weights of all links.

   Data: Table count of network flow

1. Let ii = 1;
2. while ii < table_cnt do
3. while jj < table_cnt do

4. if ii < n then
5. jj = 1;
6. ifjj < n and fiijj = 0 then
7. ifjj = I then
8. Fiijj = fmax - fiijj;
9. Wiijj = Fiijj + Piijj;
10. else
11. end if
12. jj = jj + 1;
13. else
14. end
15. ii = ii + 1;
16. else
17. return; end
18. end
19. end

3. *LEAST WEIGHT FUNCTION*: Least weight selection function that choose the path with least weight from node to sink. Algorithm for Least Weight Function is as follows.

Data: Function weight

1. Set j = 1;
2. if j < n then
3. k = 1;
4. if k < n then
5. if Wjkf = 0 then
6. Sjk = wjk
7. else
8. k = k + 1 1
9. end
10. else
11. Select node which is intermediate and has smallest Sjk
12. j = j + 1
13. end
14. else return;
15. end

OPTIMAL PLACEMENT FUNCTION: Optimal placement function that is used in case of node failure using algorithm mentioned below.

1. If neighboring nodes are connected then
2. delete failed node from WSN
3. else
4. if recovery node for given node is available
5. then replace failed node with recovery node

```
else
If recovery node is available for given node
then remove failed node and add nearer recovery node to WSN
else
k = node number which is failed
end
if k < n
then move the next node
else return;
end
end
return;
end; return
```

Enhanced OMRA finds all flows from source to sink and calls corresponding functions to find the optimum route.

3 Results and Discussions

3.1 Tools Used

- Linux OS
- NS-2 (Network Simulator) tool
- GCC compiler
- tcl-devel tool
- libX11 and Xtdevel tool.

3.2 Results of Experiments

The proposed algorithm is implemented in NS-2 and the results are given below *Simulation Environment*: Proposed OMRA is simulated in NS2. To check efficiency, proposed algorithm is compared with routing protocols such as Radio Aware (RA). Simulation parameters are given below shown in Table 1.

(1) *Simulation Results*: Parameters which are considered for testing proposed OMRA protocol are shown in Table 2.

As mentioned in tabular format of the obtained result, it is observed that the implemented system is more superior to techniques currently in vogue. The problems of data links failure is reduced to a large extent, energy conservation is improved a lot, average energy per node is only 0.0497473 which is 10–20% improved if compared to traditional methods. There are only 29 number of packets have been dropped,

Table 1 Simulation variables

Area of WSN	500×500
Total nodes in WSN	110
Initial energy	10 Jl
Noise level	0.0001
Cost per bit (CPB)	0.00005
Data reduction ratio (Delta)	0.48
Aggregation cost in WSN	1
Receiving power of nodes	0.01
Transmission power of nodes	0.02
Total simulation time	200

Table 2 Simulation results

Parameter	OMRA Algorithm	RA algorithm
Number of packets sent in WSN	620	620
Number of packets received in WSN	591	501
Packets delivery ratio in WSN (%)	95	81
WSN Control overhead	9584	9890
WSN Normalized overheads: routing	16.274	18.75
Total delay	0.044023	0.639021
Total throughput	14641.2	23464.9
WSN: Jitter	0.262275	0.45343
Number of dropped packets	29	119
Dropping ratio in WSN	4.67	19.19
Total energy consumption of WSN	4.92498	6.46532
Average energy Consumption per node	0.00797473	0.010419
Overall energy in WSN: residual	955.072	1002.435
Average energy in WSN: residual	9.65025	10.63430

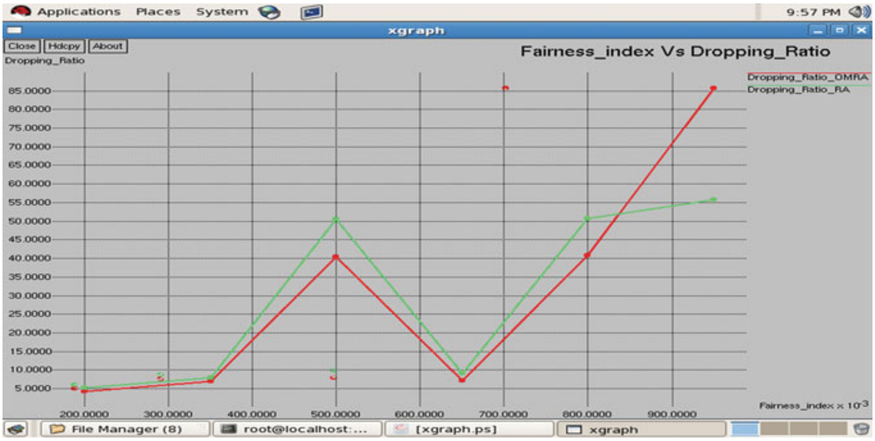


Fig. 4 Fairness index versus dropping ratio

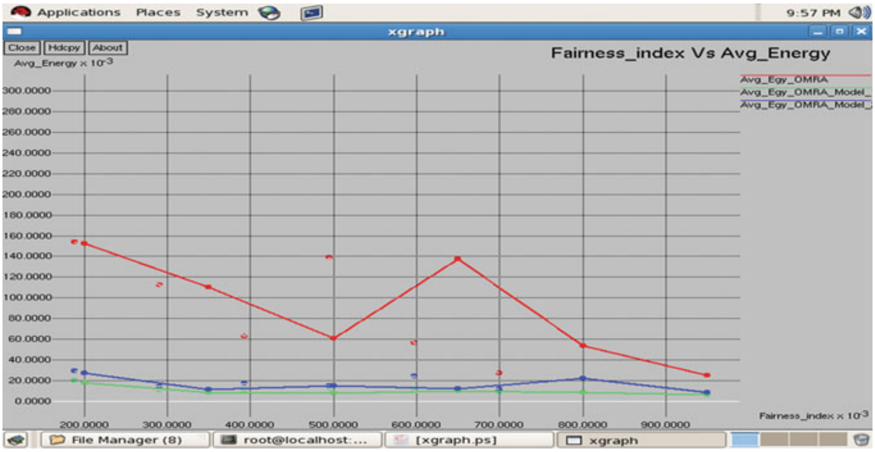


Fig. 5 Average energy versus fairness index

which is about five percent of total packets send which can be ignored considering simulation environment. Speed of communication also seen a marginal improvement over traditional technique.

Fairness index is the terminology used in WSN to determine whether applications or users are getting fair share of resources available. Figure 4 shows fairness index versus dropping ratio. From Fig. 4, it can be seen that dropping ratio in WSN (with OMRA algorithm) is less compared to dropping ratio in WSN (with existing RA algorithm). This proves the effectiveness of OMRA algorithm in terms of drooping ratio (Fig. 5).

4 Conclusion

The proposed algorithm considers the fact that network can be stable or unstable. Also, links across the nodes can be asymmetric or symmetric. Due to this approach, proposed algorithm solves instability and asymmetric link problems in WSN. For many existing algorithms, it is not possible to maintain network connection during network instability which is solved by proposed algorithm. The proposed enhanced OMRA optimizes the RA routing algorithm and has low communication complexity.

References

1. Li, J., Liu, D.: An Energy aware distributed clustering routing protocol for energy harvesting wireless sensor networks. In: 2016 IEEE/CIC International Conference on Communications in China (ICCC)
2. Rizk, R., Elhadidy, H., Nassa, H.: Optimized mobile radio aware routing algorithm for wireless sensor networks. IET Wirel. Sens. Syst. **1**(4), 206 (2011)
3. Kokare, A.J., Chavan, M.K.: Energy efficient routing protocols for wireless sensor network: a surve. In" TEICC, Bikaner, Rajstan (2012). ISBN: 978-81-923777-0-4.
4. Sharma, D., Bhondekar, A.P., Ojha, A., Shukla, A.K., Ghanshyam, C.: A traffic aware cluster head selection mechanism for hierarchical wireless sensor networks routing. In: 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (2016)
5. Sarvamangala, D.R., Raghavendra Kulkarni, V.: Multistage localization in wireless sensor networks using artificial bee colony algorithm. Commun. Comput. Inf. Sci. **776**, 451 (2017). ISSN 1865-0929, ISBN 978-981-10-6429-6
6. Krishnamachari, B., Ordonez, F.: Analysis of energy-efficient, fair routing in wireless sensor networks through non-linear optimization. In: Proceedings IEEE Semiannual Vehicular Technology Conference (VTC), USA, vol. 5, Oct 2003
7. Handziski, V., Kopke, A., Karl, H., Frank C, Drytkiewicz, W.: Improving the energy efficiency of directed diffusion using passive clustering. In: Proceedings of the First European Workshop on Wireless Sensor Networks, EWSN, Berlin, Germany, vol. 2920. LNCS
8. Kulkarni, R., Fo´lrster, A., Venayagamoorthy, G.: Computational intelligence in wireless sensor networks: a survey. IEEE Commun. Surv. Tutor. **12**(3), 1–29 (2010)
9. Yuanrong, C., Jiaheng, C.: An improved directed diffusion for wireless sensor networks. In: Proceedings of WICOM, Shanghai, China
10. Zhiyu, L., Haoshan, S.: Design of gradient and node remaining energy constrained directed diffusion routing for WSN. In: Proceedings of WICOM, Shanghai, China
11. Chen, M., Kwon T., Choi, Y.: Energy-efficient differentiated directed diffusion (EDDD) in wireless sensor networks. Comput. Commun.
12. Pyokin, Y., Jung, E., Park, Y.: A radio-aware routing algorithm for reliable directed diffusion in lossy wireless sensor networks. Sensors **9** (10) (2009)
13. Guerriero, F., Violi, A., Natalizio, E., Loscri, V., Costanzo, C.: Modeling and solving optimal placement problems in wireless sensor networks. Appl. Math. Model. **35**(1) (2010)
14. Kalpakis, K., Dasgupta, K., Namjoshi, P.: Maximum lifetime data gathering and aggregation in wireless sensor networks. Int. J. Comput. Telecommun. Netw. **42**(6), 697 (2003)
15. Intanagonwiwat, C., Govindan, R., Estrin, D.: Directed diffusion: a scalable and robust communication paradigm for sensor networks. In: Sixth Annual International Conference on MobiCom, USA

A Study on LoRaWAN for Wireless Sensor Networks

S. Subashini, R. Venkateswari and P. Mathiyalagan

Abstract Wireless sensor network plays a tremendous role in various fields such as agriculture, environmental monitoring, military applications, health care, etc. There are many challenges in designing wireless sensor network that are specific to the application under consideration. The developments of Internet of Things with interconnected physical objects improve the application space, flexibility and sophistication of wireless sensor networks. LoRaWAN is a long range wide area network which uses low power and unlicensed Lora Band for wireless communication among battery operated devices. The characteristics such as transfer rate 300 bps to 50 kbps, low power and very low duty cycle makes LoRaWAN a potential candidate for IOT applications. In this paper the possibility of implementing LoRaWAN for variety of wireless sensor network application have been analyzed.

Keywords Wireless sensor network (WSN) · Internet of things (IoT) LoRaWAN

1 Introduction

Wireless sensor network is a collection of low-powered sensor nodes to collect the data from the environment such as temperature, moisture, humidity, etc. The collected data is processed by the processing subsystem and accessed by the users remotely from the server through various communication technologies [1]. The sensor nodes depend on batteries for power supply and energy harvesting by means of solar panels in the node is also possible. There are many challenges in wireless sensor network

S. Subashini (✉)
Department of CSE, PSV College of Engineering and Technology, Krishnagiri, India
e-mail: subashinianandraj@gmail.com

R. Venkateswari
Department of ECE, PSG College of Technology, Coimbatore, India

P. Mathiyalagan
Department of CSE, Sri Ramakrishna Engineering College, Coimbatore, India

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_26

such as energy conservation, fault tolerance, security, cost, and middleware. Wireless sensor networks are applied in a wide variety of applications such as environmental monitoring, industrial monitoring, agricultural monitoring, pollution monitoring, disaster management, and military applications [2].

The growth of sensor network lead to the development of Internet of Things (IoT) Internet of Things (IoT) is a network which connects many physical objects and devices with embedded sensors, computing power and networking capabilities [3]. These objects are connected by various technologies such as RF identification (RFID) WiFi, ZigBee, and other wireless technologies. The Internet of Things (IoT) is offering solutions towards various applications such as smart city, smart agriculture, smart home, intelligent transportation and in various industries [4].

With the development of Internet of Things (IoT), the WSNs will undergo a new revolution and its applications can be more interesting and exciting. In order to integrate these applications into IoT, many challenges to be resolved such as latency, interoperability, range coverage and bandwidth. A suitable wireless technology plays an important role in solving these issues. The IoT application involves remote monitoring in different environments and has to avoid expensive installations. Many protocols like SigFox, NB-IoT and LoRa came into picture and comparison is given in [5]. LoRaWAN is a type of Low Power Wide Area Network (LPWAN) technology designed for IoT applications to address such issues [6]. It possesses various advantages like reliability, scalability, delay tolerance, interference immunity and achieves a relative long distance. The LoRaWAN will have different capabilities as per application requirements. Hence the motivation of this study is to analysis LoRaWAN in wireless sensor network point of view.

The rest of the paper is organized as follows. The Sect. 2 presents the LoRaWAN architecture, device classes, specification and energy performance. The Sect. 3 presents the applications of LoRa for various IoT based WSN applications. Finally the paper is conclusion is given in Sect. 4.

2 LoRaWAN

LoRa is a LPWAN protocol with long range and low power suitable for smart sensing technology in infrastructure based systems and industrial applications. The devices having limited energy and the traffic initiated by both the sensors and actuators are gaining advantage by LoRa due to its low power and long range [7]. LoRa uses chirp spread spectrum (CSS) modulation with varying spreading factors and bandwidth [8, 9]. The spreading factor ranges from 7 to 12 and the gateway receives multiple transmission depending on the spreading factor leads to high interference immunity. It is a new technology by Semtech Corporation operates at various ISM bands 433, 868, and 915 MHz [10]. LoRa is referred by physical layer and MAC layer where the former uses chirp spread spectrum technique and the later uses LoRaWAN for medium access control mechanism [7]. The trade-off between the power consumption and transmission range is achieved in LoRa compared to Bluetooth, Wifi and cellular networks.

2.1 LoRaWAN Architecture

LoRaWAN network is star of stars topology as shown in Fig. 1 which defines a communication protocol that works on the top of the LoRa physical layer. The components of the network are end devices, gateways, network server and application server. The end devices are equipped with sensor nodesthat are responsible for information gathering and send to the gateways by one hop mechanism. They are connected to multiple gateways which forward the data to the network servers. The network servers are responsible of filtering the duplicate packets, security, sending the acknowledgements and forwards data to the application server [8].

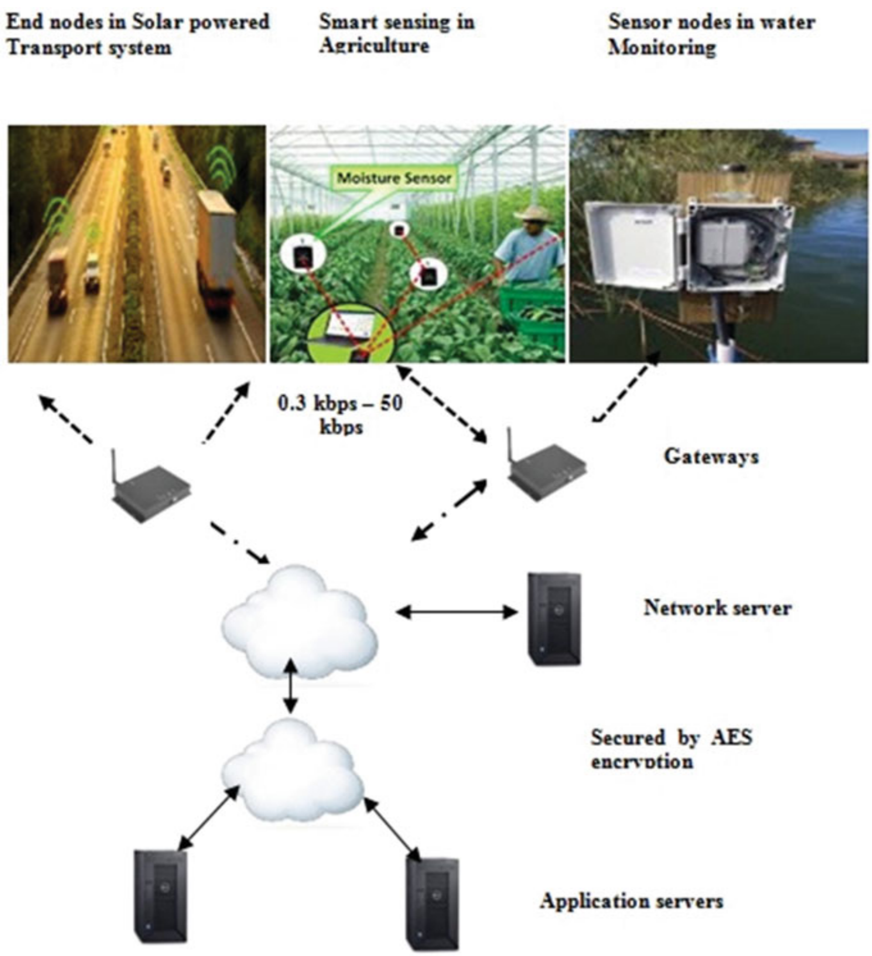


Fig. 1 LoRaWAN network architecture

2.2 Device Classes

There are three types of classes (A, B and C) in LoRaWAN according to specific application needs which are described as follows.

Class A. It allows bidirectional communication, having lowest complexity and low energy consumption. The transceivers of the end devices are in deep sleep mode for the maximum duration and wake up infrequently for data transmission [9]. The devices receive the downlink messages in their one or two receive windows after each upstream transmission in the appropriate time slots [9].
Class B. It also provides bidirectional communication and with an additional advantage of scheduled time slots to receive the data. The beacon messages are used for synchronization between the end devices and the gateway to make the network servers aware of the listening time of end devices [5–7]. The class B devices receive slots for downstream transmission without upstream transmission as in class A [11].
Class C. The class C also allows bidirectional communication with more receive slots for continuous downlink transmission except during data transmission [12]. This class of devices achieve lowest latency than class A and class B but the energy consumption is high [6].

2.3 LoRaWAN Specification

The features of LoRaWAN is analyzed by many researchers [6, 8, 13] and compared with other leading technologies. The specification of LoRaWAN is listed in Table 1.

Table 1 LoRaWAN specifications

Features	LoRaWAN
Modulation	Chirp based spectrum
Data rate	300–50 kbps
Link budget	154 dB
Interference immunity	Very high
Scalability	Yes
Transmission range	15–30 km
Battery lifetime	8–10 years
Energy efficiency	High
Bandwidth	433, 868, 915 MHz

2.4 Energy Performance of LoRa

Many researchers measure the energy consumption of the sensor node using first order radio model [14]. The energy required for transmission of data of n bits over a distance d is

$$\mathrm{E_T}(n,\ d) = \mathrm{E_{Tx}}(n) + \mathrm{E_{Tamp}}(n,\ d) \tag{1}$$

$$\mathrm{E_T}(n,\ d) = \mathrm{E_{Tx}}.n + \mathrm{E_{Tamp}}.n.d^2 \tag{2}$$

The energy spent for receiving data of n bits is

$$\mathrm{E_R}(n) = \mathrm{E_{Rx}}(n) = \mathrm{E_{Rx}}.\,n \tag{3}$$

Carles Gomez measured the energy consumption by the end devices by considering the various states of devices and the energy consumed in each state as given in Eq. 1 [12]. Average current consumption of an end device is represented by Eavg

$$\mathrm{E_{avg\_ua}} = \frac{1}{T_{int}} \sum\nolimits_{i=1}^{N} T_i.I_i, \tag{4}$$

where $\mathrm{T_{int}}$ denotes the time between two consecutive data transmissions by the end devices, N denotes the number of states of the end devices, $\mathrm{T_i}$ denotes the duration of the state i and $\mathrm{I_i}$ denotes the current consumption of nodes in state i. The current consumption is measured in various states such as wakeup, sleep, transmission, radiooff, etc.

3 Analysis of LoRaWAN for WSN Application

The following Table 2 summarizes the various IoT-based applications, their requirements and the suitability of protocols.

A framework of Google applications are used for resource management in construction site. The IoT platform enables coordination among the persons, vehicles, location to share the information among them. The information from multiple domains is merged for monitoring and controlling in terms of logistics and technological requirements. The machine to machine technology is facilitated by mobile networks for their good coverage, high throughput and low deployment cost but the high energy consumption and operational cost leads to the searching of technologies with low power and long distance like LoRa [15].

The wireless sensor network used to predict the landslides in the unstable slopes needs continuous monitoring resulting in high energy consumption. The system continuously monitors the overall shape deformation of the slopes to prevent the landslides. The IMST's iM880A low power radio module is using LoRa modulation

Table 2 LoRa for IoT based applications

Typical application of sensor networks	Requirement	Suitability of protocols	Literature survey
Water quality monitoring	Continuous monitoring, Efficient data transmission	Zigbee, Wifi, LoRa	Internet of things enabled real time water quality monitoring system [20]
Resource management at construction site	Faster monitoring and controlling	LoRa	A platform using IoT for quality management of construction [15]
Landslide prediction	Continuous monitoring	LoRa	Implementation of LoRa to predict landslides for long range communication [16]
Agriculture	Distributed measurements	Zigbee Wifi, LoRa	IoT applications in agro-industrial and environmental fields [17]
Underground sensor networks	Periodic monitoring	Wifi, Lora	IOT based underground grid system [18]
Marine environment monitoring	Periodic data collection	LoRa	Wireless Sensor Network for marine-coastal environment monitoring using unmanned aerial vehicles and buoys [21]
Smart city	Continuous monitoring	IEEE802.11p IEEE802.11n LoRa	A mobile sensor network based bus assisted smart city sensing for environmental data collection [19]

to achieve long range and high interference immunity. The combination of forward error correction techniques and spread spectrum modulation increased the range of LoRa modulation. The 868 MHz band provides high interference immunity with minimized power consumption. The LoRa covers a range of 30 km suitable for the positioning of sensor nodes and anchor node which are placed at a distance of 10–15 km [16].

The agricultural monitoring involves measurement of the physical and environmental variables using wireless sensor network and remote sensing technology. The applications of air monitoring, water monitoring and soil monitoring measures the data in a periodic and continuous manner. The continuous monitoring makes the devices to wake up continuously leads to depletion of energy. The sensed data is transmitted either between the end devices or between the end devices and gateway or a two way communication in terms of sensor and actuator network. The LPWAN technologies such as SigFox, LoRa, NB-IOT are adopted in IoT-based agricultural applications to achieve high energy conservation, long range with low cost. Many researchers analyzed various technologies and found LoRa is best suited for smart agricultural applications [17].

An under ground safety hybrid (USAH) architecture is proposed to increase the accuracy and reliability where the sensor nodes collect the data periodically combined with aperiodic sensing. The WAPUS (Wireless Access Point for Underground Safety services for Sinkhole detection) comprises of multiple network connections, two RF modules, NXP Cortex-A7 dual processor with 1 GHz DDR 3 memory and 4+ UARTports. The system involves periodic monitoring for low power consumption and whenever the threshold value increases aperiodic sensing takes place. The WAPUS system designed for sink hole detection applications increased the accuracy using the hybrid scheme [18].

In [19] P. Cruz et al. proposed a system to minimize the delay in mobile sensor networks in IoT applications. They have used sensor nodes fixed in buses to collect the data and the bus stops act as sink nodes. The sensor nodes transfer the data to the sink node which transfers to the task manager node. There are many bus stops in the city and all the bus stops cannot take part as sink node and this is formulated as Integer Linear Programming problem to reduce the network delay. The network delay is the maximum travel time between two sink nodes. An approximation algorithm using a greedy based approach is proposed to select the sink nodes in order to reduce the network delay.

The IoT-based applications needs to confront a considerable measure of difficulties in terms of interoperability, security, bandwidth, latency and other performance issues. As sensors are deployed in machines, things and various environments in this network, energy conservation is a significant task. To address these challenges and to enable the communication anywhere, anytime and by anything, LoRa WAN is more suitable.

4 Conclusion

In this paper, different applications of sensor network have been studied with LoRaWAN. This paper also gives insight into the LoRaWAN architecture, device classes and its specification. As it mostly uses unlicensed spectrum band in the sub GHz range and it requires very low power 140–160 decibels (dB) path to connect devices in miles of ranges, LoRaWAN is a promising technology for future IOT applications. Despite its advantages there are many challenges that need to be addressed to integrate this technology into a given application.

References

1. Anastasi, G., Conti, M., Francesco, M.D., Passarella, A.: Energy conservation in wireless sensor networks: a survey. Ad Hoc Netw. **7**, 537–568 (2009)
2. Yick, J., Mukherjee, B., Ghosal, D.: Wireless sensor network survey. Comput. Netw. **52**, 2292–2330 (2008)
3. Tzounis, A., Katsoulas, N., Bartzanas, T., Kittas, C.: Internet of Things in agriculture, recent advances and future challenges. Biosys. Eng. **164**, 31–48 (2017)
4. Chen, S., Xu, H., Liu, D., Hu, B., Wang, H.: A Vision of IoT: Applications, challenges, and opportunities with China perspective. IEEE Internet Things J. **1**(4) (2014)
5. Raza, U., Kulkarni, P., Sooriyabandara M.: Low power wide area networks: a survey. arXiv:1606.07360(2016)
6. Jonathan, D.C.S., Joel Rodrigues, J.P.C., Antonio Alberti, M., Solic, P., Andre Aquino, L.L.: LoRaWAN—a low power WAN protocol for internet of things: a review and opportunities. In: 2nd International Multidisciplinary Conference on Computer and Energy Science. IEEE (2017)
7. Augustin, A., Yi, J., Clausen, T., Townsley, W, M.: A study of LoRa: long range & low power networks for the internet of things. Sensors **16**, 1466 (2016)
8. Sinha, R, S., Wei, Y., Hwang, S.: A survey on LPWA technology: LoRa and NB-IoT. ICT Express, (3)14–21 (2017)
9. den Abeele, F.V., Haxhibeqiri, J., Moerman, I., Hoebeke, J.: Scalability analysis of large-scale LoRaWAN networks in ns-3. J. Latex Class Files **14**(8) (2015)
10. Wixted, A.J., Kinnaird, P., Larijani, H., Tait, A., Ahmadinia, A., Strachan, N.: Evaluation of LoRa and LoRaWAN for wireless sensor networks. IEEE Sens. (2016)
11. Adelantado, F., Vilajosana, X., Tuset-Peiro, P., Martinez, B., Melià-Seguí, J., Watteyne, T.: Understanding the limits of LoRaWAN. IEEE Commun. Mag. (2017)
12. Casals, L., Mir, B., Vidal, R., Gomez, C.: Modeling the energy performance of LoRaWAN. Sensors **17**, 2364 (2017)
13. Tzounis, A., Katsoulas, N., Bartzanas, N., Kittas, C.: Internet of Things in agriculture, recent advances and future challenges. Bio syst. Eng. **164**, 31–48 (2017)
14. Nikolidakis, S.A., Kandris, D., Vergados, D.D., Douligeris, C.: Energy efficient automated control of irrigation in agriculture by using wireless sensor networks. Comput. Electron. Agric. **113** 154–163 (2015)
15. Czart, W., Gierszala, H., Pawlinab, K., Urbańska, M.: ICT for resource management and telematics in construction sites. Proc. Eng. **208**, 27–34 (2017)
16. Romdhane, R.F., Lami, Y., Genon-Catalot, D., Fourty, N., Lagrèze, A.: Wireless Sensors Network for Landslides Prevention. IEEE (2017)
17. Talavera, J.M., Tobón, L.E., Gómez, J.A., Culman, M.A., Aranda, J.M., Parra, D.T., Quiroz, L.A., Hoyos, A., Garreta, L.E.: Review of IoT applications in agro-industrial and environmental fields. Comput. Electron. Agric. **142**, 283–297 (2017)
18. Heo, T., Hong, S., Lee, J., Lee, I.: IoT-based wireless access point for underground safety services. In: 10th international Conference on Sensor Technologies and Applications (2016)
19. Cruz, P., Couto, R.S., Costa, L.H.M.K.: An algorithm for sink positioning in bus-assisted smart city sensing. Future Gener. Comput Syst. (2017)
20. Geetha, S., Gouthami, S.: Internet of things enabled real time water quality monitoring system. Smart Water (2017)
21. Trasviña-Moreno, C.A., Blasco, R., Marco, A., Casas, R.A.: Unmanned Aerial Vehicle Based Wireless Sensor Network for Marine-Coastal Environment Monitoring. Sensors, 17, 460 (2017)

Avalanche Effect Based Vertical Handoff System for Wireless Communication

G. U. Mali and D. K. Gautam

Abstract Nowadays, due to increasing usage of the wireless technologies the movable node infrastructure always inviting lots of threats to worsen the network. Among these network failures is biggest and disturbing factors. Like a boon to this problem vertical handoff technology is acting like an effective approach in a wireless network. So many of the methodologies are introduced to handle vertical handoff more efficiently, but all are having one or another problem in the process. The proposed paper puts forwards an idea of vertical handoff situation awareness to minimize decision time as compared to other methods by comparing the hash keys at the pool manager of a wireless network pool with another network pool for successful handoff. Therefore, definitely our method enhanced the overall performance by taking a decision in a short time compared to vertical handoff system suing fuzzy logic.

Keywords Pool tile based vertical handoff (PTVHO) · Received signal strength (RSS) and mobile ad hoc network (MANET)

1 Introduction

In a heterogeneous wireless communication network, a mobility management system is used to transfer the data between nodes, in which handover is the main factor to maintain between space to base station. When a mobile node is moving towards another location from its current location, there is a need to change connection because of that role of handoff is an important. Currently, in the world normally all

G. U. Mali (✉)
Department of Electronics Engineering and Technology, North Maharashtra University,
Jalgaon, Maharashtra, India
e-mail: ganeshumali@gmail.com

D. K. Gautam
Head of Electronics Department, Department of Electronics Engineering and Technology, North Maharashtra University,
Jalgaon, Maharashtra, India
e-mail: dkgautam_nmu@hotmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_27

the digital mobile equipment is moving so mobility is an important issue, therefore, to overcome this problem there is a method known as handoff. If a handoff executed within same technology, called horizontal handoff and executed within different technology called vertical handoff [1].

Wireless Network is a wired free network set up among network nodes and uses radio signal frequency for communications between network devices. The setup of wireless (also called WiFi) network is simple and it avoids costly cable connection between different network nodes. Radio waves are use of wireless network for connections such as cell phone, laptop, etc. Wireless networks main types named as WAN, MAN, PAN, and WLAN. The two main components used to setup wireless connection between devices are wireless routers and wireless clients. Operating modes in wireless networks are of two types first is Infrastructure mode in that device is associated with adapters of a wireless network to a set up a wired network with the support of wireless routers and the second is Ad hoc mode where two wireless devices are directly connected without the help of a wireless router.

At present lot many methods of vertical handoff are proposed. Normally algorithm uses only RS as a parameter. (MADM which consider the many attributes to select a good network of available networks. In [2] all methods are compared like for an Ideal Solution proposed order preference similarity [4], Simple Additive Weighting [3], Grey Relational Analysis [5] and Analytic Hierarchy Process [6]. But the problem of handoff not solved any such a MADM Methods so this article proposed a fuzzy based vertical handoff system to overcome all the drawbacks of vertical handoff algorithms.

To improve the handoff probability, handoff solution based GPS proposed by Debabrata sarddar for wireless system. Direction of mobile terminal velocity (MT) decided to use GPS. By using this they ensured efficient handoff. In this method actually they stored a different time interval angles into the memory and compared with specific interval. therefore for accurate result they required to stored a huge data. So it requires a memory with high capacity [7].

From that mainly decision of handoff is taken by the mobile user velocity with the network coverage area and find the critical speed for the particular network coverage area as per application at the time of handoff is essential proposed by Jain. Here they used mobile velocity and coverage range for vertical handoff decision. This decision improves the case as well as reducing no. Of unwanted hands-off. But this proposed method requires to be considered more parameters for proper vertical handoff decision [8].

Goyal [9] proposed a model based on dynamic decision proposed by for decision of vertical handoff in wireless networks (ADDMVHO). This system considered dynamic factors like the RSS, mobile velocity and static factor too in the right vertical handoff decision to select the strong network at best time among the available network. So in that method a one handoff management center (HMC) also taken input from base stations (BS) and network interfaces And after analysis of this valuable information it will go for vertical hand off decisions. The algorithm used in this method has a different phases. The first phase is Priority phase where all the ineligible and unwanted removed from all available candidate networks. Then in next Normal

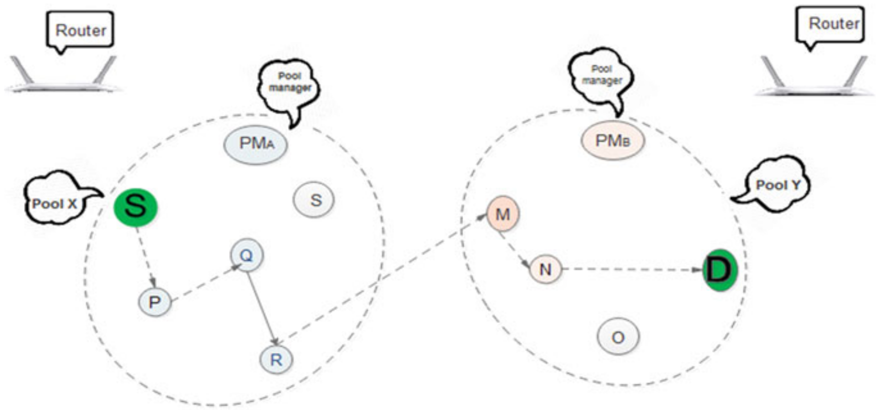


Fig. 1 Overview of the proposed model

Phase is executed for user specific preferences concern to network interface usage. In last Decision Phase is used to choose best network by taking vertical handoff decision.

Latency is also a important issue therefore latency reducing purpose a new vertical handoff system proposed by Rana [10]. When wireless node moves from its current location to another location, it updates with a new address from the subnet of a particular network after analysis of router advertisement given by the access router. Then wireless node updates itself with new address. Then it start to verify this address whether any duplicate address is there in same radio range. For that purpose it performs DAD(Duplicate Address Detection) process. If duplicate address is found then again mobile node gets new address, updates it and perform the DAD operation till it will get original address. It takes once it receives a new address for vertical handoff. But up to certain level, it minimizes the latency so to reduce the same and for the proper vertical handoff decision needs to consider more parameters too. The rest of the paper is organized as follows. Section 2 presents the design of our approach. Section 3 presents the result and discussion Sect. 4 provides the conclusion of this research paper.

2 Proposed Mehodlogy

Proposed methodology of vertical handoff situation awareness can be seen in the below Fig. 1.

Step 1: initially in our system the source evolves the shortest path that is discussed in our past research work.

Step 2: for better understanding considers shortest path is available to be {S, P, Q, R, M, N, O, D}. In that Nodes {S, P, Q, R} belongs to pool X and Nodes

{M, N, O, D} belongs to pool Y. As the pool is established and source node is started with transferring the data using 2PC protocol [11].

Step 3: Here in this step all nodes on its data hopping it creates an instance path belongs to the respective pool and keep creating a hash key using MD5 hashing algorithm on each hop and sends back to the respective pool manager to keep an eye for pool changing. So, according to this all the nodes of Pool X and Pool Y are creating a hash key for the routing path.

Step 4: Avalanche effect- Here in this phase of the proposed model pool manger of Pool X keep recording the hash key received by its nodes {S, P, Q, R}. And comparing each hops hash keys with the previous one. As the node R transfers the data it creates the hash key for its instance destination node that is M 's pool routing path which is {M, N, D} and sends it back to pool manager of pool X.

And now the hash key of the node R and hash key of the Node Q is compared by the pool manager of the pool X. The small differences in hash keys create an avalanche effect in the difference. This difference indicates the change of the data transfer in between the pool though nodes X and Y. This eventually triggers the vertical handoff decision making. This can be shown in below mentioned algorithm 1 [14].

ALGORITHM 1: AVALANCHE EFFECT

// Input: Sender Data D
Destination Node Dn
// Output: Vertical handoff Decision through Avalanche Effect
Step 0: Start
Step 1: Set vertical_handoff_ Flag=false
Step 2: Add Pool managers M1 into pool P1, and so on Mn to Pn
Step 3: Add node Nn to pool Pn
Step 4: Activate all pool managers Mn
Step 5: Select Data D by source node Sn
Step 6: Select Destination node Dn
Step 7: Identify the shortest path Pth
Step 8: WHILE D ₡ to Dn
Step 9: For each Hop H
Step 10: Routing path hash key HC and HP
Step 11: HC $\rightarrow$ Pi
Step 12: HP $\rightarrow$Pi
Step 13: IF HC $\neq$ HP
Step 14: Set vertical_handoff_ Flag=false
Step 15: End IF
Step 16: End For
Step 17: End While
Step 18: return flag
Step 19: Stop

Table 1 Comparison of vertical handoff time

Sr. No.	Fuzzy logic based vertical handoff descision (time in milliseconds)	Avalanche effect based vertical handoff descision (time in milliseconds)
1	25	17
2	40	34
3	44	41
4	60	50
5	60	54
6	70	61
7	80	65
8	82	66
9	82	71
10	84	71

3 Result and Discussion

The proposed system of vertical handoff situation awareness system is deployed in real-time scenario by using 20 computers of windows and each device processor is core i3 processor and 4 GB of primary memory with Dlink 2 antenna router. And proposed system is developed on the Java platform by using Netbeans 8.0 as IDE.

For evaluation decision-making speed of the proposed technique is compared with the methodology of [12] for vertical handoff. The accumulated results are tabulated in the below Table 1 for the time parameter by comparing first 10 hands-off time with that of the traditional fuzzy logic method [14].

In the above plot in Fig. 2 red worm belongs to our proposed system of vertical handoff using avalanche effect which crawling below of blue worm which belongs to vertical handoff decision on fuzzy logic based. By observing above plot it is clear that our proposed system takes a very less time for decision than system design using fuzzy logic. This is because of using the pool tile (PTVHO) method along with an avalanche effect which eventually takes off the burden of handling handoff of nodes [14].

When it is compared with the movement—aware-based vertical handoff [13]. Below plot in Fig. 3 clearly indicated that our method PTVHO reduces no. of handoff [15].

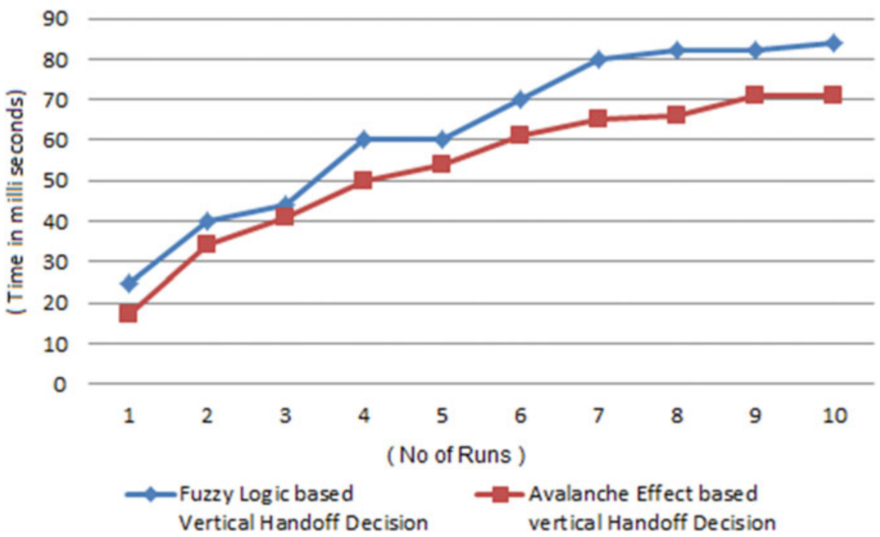


Fig. 2 Comparison of vertical handoff tim

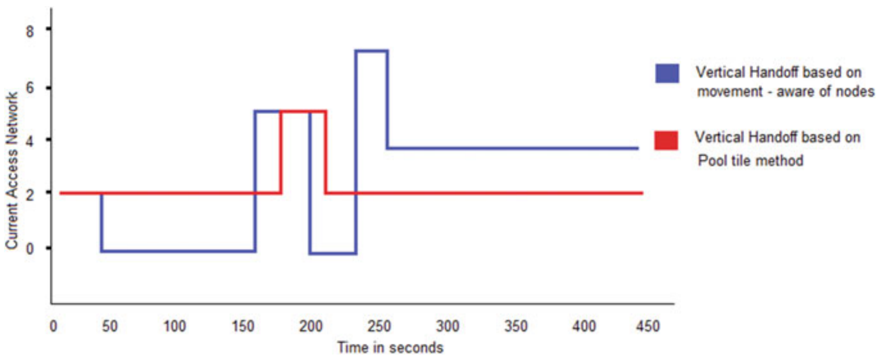


Fig. 3 Comparison for number of Handoffs

4 Conclusion

This paper proposes an idea of collecting the routing information of defined attribute for the instance network pool and creates hash key using MD5 Algorithm. And this info send to pool manager for each hop. As the network pool changes it affects the hash key of attributes so an avalanche effect is happening, which triggers the handoff process during the routing process. Means it takes a less decision time also in the same way it reduces the number of vertical handoff too.

In the coming edition of our research this idea will be deployed and compared with many other methodologies that eventually triggers the vertical handoff process.

References

1. Gu, E., Song, M., Zhang, Y., Wang, L., Song, J.: Novel network selection mechanism using AHP and enhanced GA. In: 7th Annual Communication Networks and Services Research Conference (CNSR), pp, 397–401 (2009)
2. Savitha, K., Chandrasekar, E.: Vertical handover decision schemes using SAW and WPM for network selection in heterogeneous wireless networks. Global J. Comput. Sci. Technol. **II**(9) (2011)
3. Liu, S., Pan, S., Xu, M.: An improved TOPSIS vertical handoff algorithm for heterogeneous wireless networks. In: IEEE International Conference on Communication Technology (ICCT), pp, 750–754, Nov 2010
4. Savitha, K., Chandrasekar, E.: Grey relation analysis for heterogeneous wireless networks. Eur. J. Sci. Res. **54**(4), 560–568 (2011). ISSN 1450–216X
5. Radhika, K., Venugopal Reddy, A.: AHP and group decision making for access network selection in multi-homed mobile terminals. Int. J. Comput. Sci. Eng. (UCSE) **3**(10) (2011). ISSN: 0975-3397
6. Sharma1, M., Khola, R.K.: Fuzzy logic based vertical handover decision system. Int. J. Ad Hoc Sens. Ubiquitous Comput. (IJASUC) **3**(4) (2012)
7. Sarddar, D., Mandel, K., Jana, T., Biswas, U., Naskar, M.K.: A handoff technique to reduce false-handoff probability in next generation wireless networks. (IJCSE) Int. J. Comput. Sci. Eng. **02**(03), 630–634 (2010)
8. Jain, A., Tokekar, S.: Optimization of vertical handoff In UMTS and WLAN heterogeneous networks. In: IEEE International Conference on Emerging Trends in Communication, Control, Signal Processing & Computing Applications (C2SPCA) (2013)
9. Goyal, P., Saxena, S.K.: A Dynamic Decision Model for Vertical Handoffsacross Heterogeneous Wireless Networks. World Acad. Sci. Eng. Technol. WASET.ORG, **677**(41), 676–682 (2008)
10. Rana, A., Kaur, K., Aggarwal, A.: Wireless Network (3G &4G). Department of Computer Science, AIMT, Ambala, (Kurukshetra)
11. Jain, A., Tokekar, S.: Optimization of vertical handoff in UMTS and WLAN heterogeneous networks. In: IEEE International Conference on Emerging Trends in Communication, Control, Signal Processing & Computing Applications (C2SPCA) (2013)
12. Kulkarni, M.P., Bale, V.S.: Vertical handoff decision algorithm to reduce unnecessary handoff for heterogeneous wireless networks. (IJCSIT) Int. J. Comput. Sci. Inf. Technol. **6**(6), 5033–5036 (2015)
13. Xiaona, Q., Qing, L.: A vertical handoff decision algorithm based on the movement-aware of MobileNodes. In: Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (2014)
14. Manli, G.U., Gautam, D.K.: Vertical hand off situation awarness by avalanche effect. In: Communications on Applied Electronics (CAE), Newyork USA, Vol. **7**(9), (2017) ISSN:2394–4714
15. Manli, G.u.: An extensive proposal for vertical handoff technique in WLAN. In: Int. J. Adv. Engg Res. Sci (IJAERS). **4**(2), (2017) (ISSN 2349–6495)

Energy-Aware Approach for Routing Protocol by Using Centralized Control Clustering Algorithm in Wireless Sensor Networks

Nada Al-Humidi and Girish V. Chowdhary

Abstract Routing in wireless sensor networks (WSNs) has a primary task for data transfer from source to the sink. Due to restricted battery power of the sensor nodes, there is a necessity to take in consideration while designing a routing protocol in WSNs the power saving of sensor nodes. Several routing protocols employing hierarchical-based clustering technique have been proposed for WSNs, however most of them still have such challenges which can be represented in minimizing the energy consumption and maximizing the network lifetime, simultaneously. In this paper, an improved method EACCC is proposed by extending the centralized clustering technique in order to achieve higher efficiency for energy, longer lifespan of network and network scalability. The performance of EACCC is evaluated and justified through extensive analysis, analytical proof, comparison, and implementation. The results show that the proposed method is highly efficient and effective in term of balancing the consumption of energy and prolonging network lifetime.

Keywords WSNs · Routing protocol · Centralized algorithm · Network lifetime

1 Introduction

Currently, WSNs have become the furthermost exciting networking technologies. With restricted power ability, the sensed collected data is offered to the sink. "WSNs consist of a huge number of low-cost, restricted power, and multifunctional wireless sensor nodes with sensing computation capabilities and wireless communications. These sensors communicate via a wireless medium within the short distance and collaborate to accomplish a common task, such as environment monitoring, and industrial process control" [1]. Because of the limitations of network resources such

N. Al-Humidi (✉) · G. V. Chowdhary
School of Computational Sciences, S.R.T.M University, Nanded, India
e-mail: alhumidinada@gmail.com

G. V. Chowdhary
e-mail: girish.chowdhary@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_28

as energy, storage, bandwidth, several network constraints and requirements, the routing protocols designing for WSNs is challenging. Generally, routing protocols in WSNs are classified in two categories: proactive and reactive protocols [2]. Proactive routing protocols keep track of routes to all destinations in routing tables whereas reactive protocols acquire routes on demand and avoid saving information about the network topology. Depending on the structure of the network routing protocols may be classified into three categories: location-based routing, hierarchical-based routing, and flat-based routing [3]. Due to higher energy efficiency, lower data retransmission and network scalability, the routing protocols hierarchical-based are the most efficient protocols in WSNs. In such a protocol, the whole network is grouped into clusters where every cluster has a leader called as cluster head (CH) and used for aggregation of data and data transmission, and for data sensing other sensor nodes (non-CH) are used. The main challenges with the clustering algorithms are the selection of the CH and managing the clusters, several routing protocols based clustering have been proposed as a result of these challenges. The rest of the paper is organized as follows. Section 2 elaborates the previous research works and in the Sect. 3, the preliminary notations are defined. The proposed method is described in Sect. 4 and Sect. 5 shows the results and performance analysis. In Sect. 6 the conclusion is presented.

2 Related Work

Heinzelman et al. [4] designed the LEACH (Low Energy Adaptive Clustering Hierarchy) protocol. The operation in LEACH is divided into different rounds, in which every round run two phases: setup and steady-state. Clusters are formed and one node is selected to be a cluster head (CH) in every cluster in setup phase. Every sensor node is generated a random number between 0 and 1(r), r is compared with the threshold value $T(n)$ in such a way that the sensor node becomes a CH in the current round if r is less than $T(n)$, otherwise, it will be a member node. The CH is selected according to the following probability:

$$T(n) = \begin{cases} \frac{p}{1-p*(r\ mod\ 1/p)} & , if\ n \in G \\ 0 & , otherwise \end{cases}$$

After selecting the CHs, a message is broadcasted to all sensor nodes by the CHs. Based on the strength of the received signal the sensor nodes will be decided in which CH will be joined and sent a message for joining to the selected CH. Every CH sets up TDMA schedules for all member nodes of its cluster. In steady phase, the cluster members transmit their data to CH according to their TDMA schedule, CH aggregate the data then transmit it to the base station. However battery depletion is avoided and communication between sensor nodes and the base station is less in LEACH, but the CH selection has wasted the energy during setup phase, and there is no guarantee for CH distribution. Lindsey et al. [5] presented a PEGASIS (Power-Efficient GAthering in Sensor Information Systems) protocol, which is an improvement of the LEACH protocol. In every cluster, PEGASIS forms chains from sensor nodes, in which each sensor node transmits and receives data only from its neighbor and only one of these sensor nodes in the chain is elected to transmit data to the sink. The collected data transfer from one node to another, aggregated and lastly is sent to the sink. PEGASIS protocol avoids so much clustering, but the chain shaping overhead is introduced, and it leads to the problem of a packet delay. Manjeshwar et al. [6] design TEEN (Threshold sensitive Energy-Efficient sensor Network) protocol for reactive networks. If the sink has interesting attributes, then an event will report to the sink. A trade-off between the energy consumption applications and accuracy is provided. It is appropriate for real-time applications, but it is not appropriate for usual applications of data gathering. Heinzelman et al. [7] proposed LEACH Centralized protocol (LEACH-C). In this protocol, a centralized algorithm is run at the base station to select the CHs based on their energy information, and then it broadcasts a message to all sensor nodes to inform them about the CHs ID. LEACH-C is forming better balanced clusters, but it is not robust and relatively high overhead. Bakaraniya et al. [8] proposed K-LEACH (Kmedoids-LEACH)protocol. K-medoids algorithm is used for forming the clusters, and Euclidean distance is used for selecting the cluster head. All nodes in K-LEACH are homogeneous, as the clusters are formed only in the first round there is no improvement in the network lifetime. Arumugam et al. [9] proposed EE-LEACH (Energy-Efficient LEACH) protocol for data gathering. In EE-LEACH protocol, for each cluster the cluster head is selected to optimize the resource utilization and reduce the energy consumption. So, the nodes with the highest residual energy will be selected to forward the data to sink. EE-LEACH provide better ratio of packet delivery, but it is lack to provide the integrity of data. Sindhwani et al. [10] proposed V-LEACH protocol. In V-LEACH a vice-CH is selected besides having a CH in the cluster that can take the role of the CH, when the CH die. So the cluster nodes' data will always send to the sink, but it increases setup phase, and reserve a node in each round. Alnawafa et al. [11] proposed a multi-hop technique (MHT-LEACH) as an improvement of LEACH, based on LEACH protocol the CHs is elected then based on threshold distance value all CHs are divided into two groups, external group and internal group. The internal CHs transmit their data to BS directly. While each CH in the external group establishes a special routing table for choosing the next hop to the BS. In [12] Alnawafa et al. proposed an improvement for MHT-LEACH which is refer as IMHT-LEACH, instead of dividing the CHs into two groups, all CHs are distributed into a number of levels. Data is transferred among the CHs from the upper levels towards the lower levels until it reaches the BS. However, most of the existing routing algorithms have a problem with selecting of the optimal cluster head, performing the clusters and requiring high processing. As a result, there is a dire need to design an algorithm to be a proper for WSNs in terms of selection an optimal cluster head, and performing less processing. In this paper, a proposed method Energy-Aware Centralized Control Clustering (EACCC) is introduced with the objective of reducing the energy consumption average and enhancing the lifetime of network. The centralized control clustering technique is used for the selecting of the CH in the proposed method.

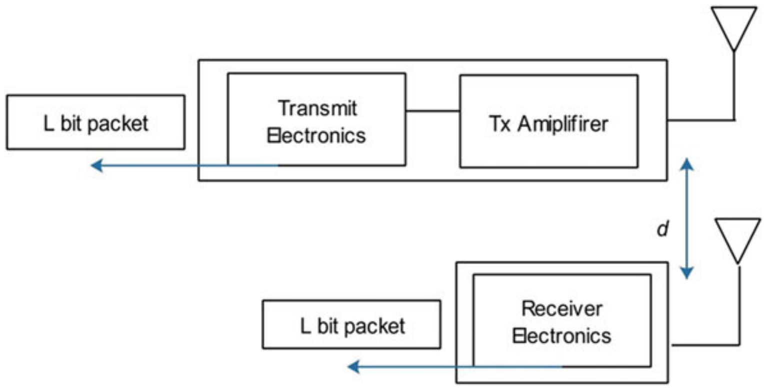


Fig. 1 First-order radio energy model

3 Preliminaries

3.1 First-Order Radio Energy Model

Many suggestions about the radio characteristics, including the effect of the performance of different protocols and energy spent in transmit and receive modes. Transmitter and receiver based on this model are shown in Fig. 1. For an individual sensor node, energy dissipation depends on the amount of data to be transmitted, transmissions number, receiving number and distance between transmitter and receiver [13, 14]. In the Fig. 1 the distance between the sender and the receiver is presented as (d) and L is the number of bits per packet transmission.

Electronics energy consumption for data transmitting ($E_{Tx}(L)$) and electronics energy consumption for receiving the data $E_{Rx}(L)$ are the same and it is given by formula (1) as follows:

$$E_{Tx}(L) = E_{Rx}(L) = E_{elec} * L \tag{1}$$

where E_{elec} is the energy spent per bit to run the transmitter or the receiver circuit. For transmit L-bit packet within distance d between any two sensor nodes, the transmission energy cost is calculated by the formula (2) as follows:

$$E_{Tx}(L, d) = E_{Txelec} * (L) + E_{Txamp}(L, d) = E_{elec} * L + E_{amp}(L, d) \tag{2}$$

where E_{amp} the amplifier of energy consumption, it can be also expressed in the terms of ε_{fs} or ε_{mp}. The ε_{fs} correspond to the free space model (*fs*), which is used when the distance between the source and the destination is less than d_0, and the $\varepsilon_{\mathrm{mp}}$ refers to the multipath model (*mp*) which is used when the distance between the source and destination is equal or greater than d_0. Where d_0 is the threshold distance. So the formula (2) can redefine as in formula (3) as:

$$E_{Tx}(L,d) = \begin{cases} (E_{elec} * L) + \left(\varepsilon_{fs} * L * d^2\right), & if\ d \le d_0 \\ (E_{elec} * L) + \left(\varepsilon_{mp} * L * d^4\right), & if\ d > d_0 \end{cases} \quad (3)$$

4 Proposed Method

The objective of the work is to propose an improved routing technique EACCC used for cluster heads selection and forming appropriate clustering so as to avoid the problem of random selecting of the CHs, guarantee that the CHs have enough energy to transmit data to the base station (BS), and offer uniformity distribution for CHs through the network area. The EACCC aims to reduce the spent of total energy in the network and prolong the network lifetime by maximizing the number of alive sensor nodes and reducing the data to be transmitted through the technique of data aggregation. The process of the EACCC is depending on the centralized control cluster algorithm which is implemented at the BS [11]. At the beginning of the EACCC, all sensor nodes send their information about residual energy and location to the BS, BS will divide the network area into four regions, and depending on information which is sent by sensor nodes the BS determines in which region the sensor nodes will be. The method runs in rounds, in which every round starts with a step of cluster head selection at the cluster head will be chosen, then the cluster forming step at which clusters are formed. Last step is the transmission of data to the BS.

4.1 The Suggested Network Model

Besides the first-order radio energy model, assumptions for the network model have to be listed in order to perform the proposed method.

- All sensor nodes are distributed randomly in a 2-Dimensional network field, energy constrained, stationary, and aware of their geographical locations and residual energy.
- A static BS is situated either inside or outside the sensor network area with unlimited energy supply.
- The sensor nodes are cluster head or cluster member node.
- The sensor nodes have the capability of power control to change their power of transmitted.
- Sensor nodes can be either homogeneous or heterogeneous.

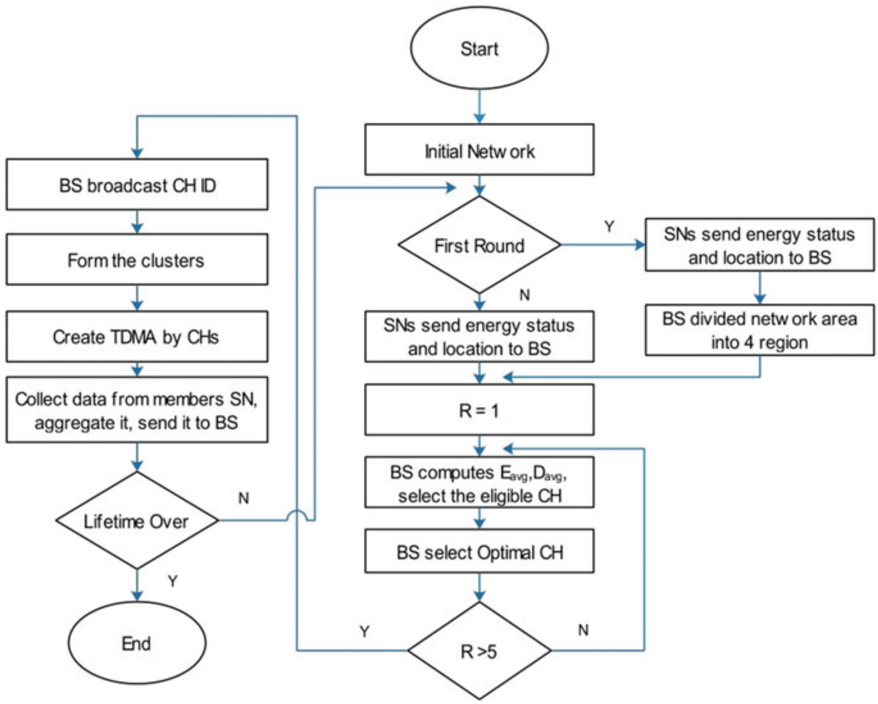


Fig. 2 EACCC flowchart

4.2 Method

The algorithm of the proposed method EACCC comprises two steps. At the first step, the network area is divided into regions based on the information sent by sensor nodes. The second step consists of three phases: selection of cluster head, cluster forming, and data transmission phase as explained below. The EACCC's flow chart is presented in Fig. 2.

First Step. At the beginning, all sensor nodes send their current energy status (how much energy is remaining in the sensor node) and locations to the BS. In order to offer uniform distribution of CHs through the network i.e. reduce isolated nodes number, the BS divides the network area into four regions (r_1, r_2, r_3, r_4), in such a way if the network field's size is $M * M$, then the regions' size will be as following formulas (4)–(7).

$$r_1 = (0 \to M/2, 0 \to M/2) \tag{4}$$

$$r_2 = (M/2 \to M, 0 \to M/2) \tag{5}$$

$$r_3 = (0 \to M/2, M/2 \to M) \tag{6}$$

$$r_4 = (M/2 \to M, M/2 \to M) \tag{7}$$

Second Step. The second step will operate in rounds, in which every round run three phases, which is cluster head selection, cluster forming, and data transmission. These phases are described as following:

Cluster Head Selection. All sensor nodes (SNs) will send their current residual energy status and locations to the BS. Suppose N is the SN number, and n is the SN number in every region which be defined in the first step, the BS computes the energy average as defined in formula (8) and the distance average from SN to BS as defined in formula (9) based on the information sent by SNs.

$$E_{avg} = \sum_{i=1}^{n} E(SN_i)/n \tag{8}$$

$$D_{avg} = \sum_{i=1}^{n} d(SN_i, BS)/n, \tag{9}$$

where $E(SN_i)$ is the sensor node's residual energy. For every region, to ensure that only SN with a high level of energy and nearer to the BS are selected as CHs for this round, the SNs with an energy level above the average of energy ($E(SN) \geq E_{avg}$) and with the distance less than the average of distance to the BS $\left(d(SN, BS) \leq D_{avg}\right)$ are qualified to be a CH candidate. Among the eligible cluster head candidate (K), the optimal No. of cluster heads is calculated by the formula (10) as in [15], and the CHs will be chosen if the eligible CH satisfy the formula (11) and (12) in every region.

$$K_{opt(r_i)} = \sqrt{\frac{n}{2\pi}} * \frac{2}{d_{toBS}} \tag{10}$$

$$max_{i=1\rightarrow K}(E(CH_i)) \tag{11}$$

$$min_{i=1\rightarrow K} d(CH_i, BS) \tag{12}$$

$$d_{toBS} = \sum_{i=1}^{K} d(CH_i, BS)/K \tag{13}$$

After selecting the CHs in every region, the number of the CHs (CH.N) in a whole network will be calculated as in formula (14). Then the BS will broadcast the CH nodes ID.

$$CH.N = K_{opt(r_1)} + K_{opt(r_2)} + K_{opt(r_3)} + K_{opt(r_4)} \tag{14}$$

Cluster Forming. During this phase, all sensor nodes should keep their receivers on. By using the CSMA MAC protocol, every CH elected in this round broadcasts an advertisement message to all the sensor nodes. Based on the received signal strength (RSS) of the message, the non-CH node decides in which CH it will belong. Once the sensor node decided, it must inform the selected CH that it will be a member of your cluster by transmitting this information to the CH. After receiving all the messages from the sensor nodes which are included in the cluster, the CH creates TDMA schedule to tell every sensor node member in the cluster in which time it can transmit data, and then broadcast back this schedule to the sensor nodes member in the cluster.

Data Transmission. Assuming sensor nodes sense data, and need to send these data. During their allocated transmission time, they can send these data to the CH. In order to minimize energy dissipation in non-CH sensor nodes, the radio of these sensor nodes turns off until the node's transmission time is allocated. Once CH receives all the data sent by non-CH sensor nodes, it performs the functions of signal processing to compress these data into a single signal, and send these data to the BS.

4.3 Analytical Proof

Challenge 1. Select cluster head.
Issue. If we use the distributed algorithm which uses the random selection of CH, then CH may have not enough energy to reach to the BS. Therefore, the data which was sent by the member nodes to the CH in that cluster will be lost.

Proposition *Use the centralized algorithm to select the CH.*

Proof In centralized algorithm, BS will calculate the sensor node's residual energy average and average of distance from sensor nodes to the BS for every region based on information which is sent by sensor nodes $E_{avg} = \sum_{i=1}^{n} E(SN_i)/n$, $D_{avg} = \sum_{i=1}^{n} d(SN_i, BS)/n$. The sensor node which satisfies: $E(SN_i) \geq E_{avg}$, and $d(SN_i, BS) \leq D_{avg}$ will be elected as a CH. Among the eligible selected cluster head nodes, we select an optimal CHs in order to guarantee that the selected CH have enough energy to reach to the BS, and energy consumption will be less as the distance from the selected CH to BS is less, so for each sensor node which satisfies the: $\max_{i=1\to K}(E(SN_i))$, $\min_{i=1\to K} d(E(SN_i), BS)$, will be elected as a CH in this round.

Challenge 2. Number of CHs.
Issue. A large enumeration of the selected CH will produce a less number of member nodes. It also will reduce the efficiency of the routing protocol. At the same time, the data collection reliability will be affected.

Proposition *An optimal number of the CH.*

Proof Let K is the CH's number in the region r, then optimal number of CH depends on how many sensor nodes in the region, what is the dimensions of the sensors field, and what is the average distance from a CH to the BS. It will be computed from the formula (10). Suppose if we have 500 sensor nodes distributed randomly in network field, number of sensor nodes in region r1 is 150, and average of distance from CHs to BS is 0.765 then the optimal number of CH in r1 is:

$$K_{opt(r_1)} = \sqrt{150/2\pi} * 2/0.765 \cong 13$$

Table 1 Parameters of simulation

Parameter	Value
Sensor field	25×25 m^2
No. of sensor nodes	500
BS position	(0,0)
Initial energy	Homogenous 0.1 J Heterogeneous 0.1–0.3 J
Transmission range	5 m
E_{elec}	50nj/bit
ε_{mp}	0.02 pj/bit
Data packet size	2000 bit
Control packet	24bit

Challenge 3. Isolated sensor node.
Issue. The random selection of the CHs may result in isolated sensor nodes which may be located far away from selected CHs and communicated with the BS by spending too much energy.

Proposition *Divide the network area into regions.*

Proof Suppose we have N sensor nodes distributed randomly in network filed (M * M), the BS divides the network area into four regions (r_1, r_2, r_3, r_4), the size of regions will be as in formula (4–7). Based on information sent by sensor nodes to the BS, the BS will determine in which region the sensor nodes will be. The proposed method will be done in every region, in such a way that every sensor node will be either a CH or member node so we guarantee that the uniform selected CH among the whole network area. In addition, to form a cluster the CH will broadcast an advertisement message to all sensor nodes in the region. As the non-CH node decision for joining based on the RSS of the advertisement message, so we ensure that every sensor node will receive the advertisement message.

5 Result and Performance Analysis

EACCC, LEACH [4] and EE-LEACH [14] clustering algorithm for WSNs are simulated to prove the efficiency of the proposed algorithm. There are 500 sensor nodes deployed randomly in the 25×25 m^2 with the transmission range 5 m and 2000 bit of the data packet size. The parameters used in the simulation are presented in Table 1.

Table 2 Number of died nodes versus rounds

Routing protocol	Homogeneous		Heterogeneous	
	1st node died	50% node died	1st node died	50% node died
LEACH	115	257	117	543
EE-LEACH	134	284	131	583
EACCC	190	358	318	627

5.1 *Performance Evaluation on the Basis of the Sensor Nodes' Energy*

To evaluate the EACCC performance, the initial energy of the sensor node is taken through homogeneous and heterogeneous. 0.1 J is taken as initial node energy for the homogenous nodes and from 0.1 to 0.3 J as initial node energy for the heterogeneous nodes. The comparison between the LEACH, EE-LEACH, and EACCC is taken when the first node dies and 50% of sensor nodes die as the evaluation criterion for the simulation with respect to the type of the sensor nodes initial energy which is homogenous or heterogeneous. Table 2 shows for the homogenous nodes, the nodes are started to die in LEACH at round 115 whereas in EE-LEACH first node died in the round 134, and in EACCC first node died at round 190. At round 257 the 50% of sensor nodes died in the LEACH, and at round 284 the 50% of sensor nodes died, whereas in EACCC 50% of sensor nodes died at round 358.

For the heterogeneous nodes the first node died at round 117 in the LEACH, at round 131 first node died in the EE-LEACH, and at round 318 first node died in EACCC. At round 543, the 50% of sensor nodes died in the LEACH, at round 583 50% of sensor nodes in the EE-LEACH, whereas in EACCC 50% of sensor nodes died at round 627. So it has been cleared that the number of alive sensor nodes is increased in EACCC, therefore the network lifetime is prolonged.

5.2 *Performance Evaluation on the Basis of Efficiency Metrics*

The performance evaluation shows the efficiency and performance of the EACCC over the LEACH, and EE-LEACH through the following metrics: network lifetime, total network energy, end-to-end time delay, and number of failed nodes.
Network Lifetime. It can be defined as "the time passing from the initial deployment of the network to the moment of the connectivity reaching the specified threshold" [16]. So in EACCC, the network lifetime is calculated as the length of time from deployment of the network until either the first node or 50% nodes died, i.e., the rounds number. Figures 3 and 4 show that the lifetime of network in EACCC is

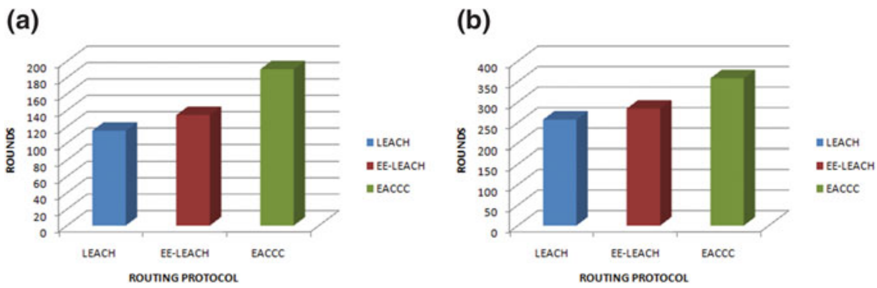


Fig. 3 Network lifetime for homogeneous nodes. **a** First node died. **b** 50% node died

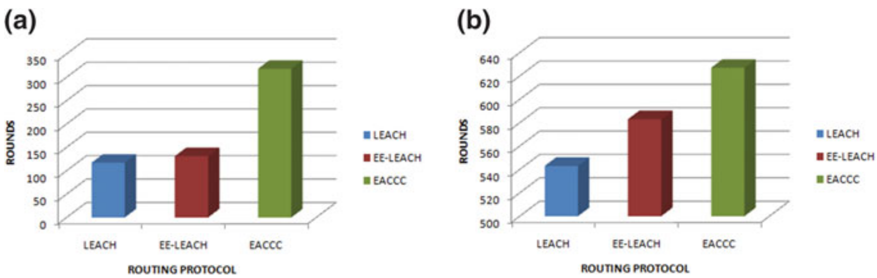


Fig. 4 Network lifetime for heterogeneous nodes. **a** First node died. **b** 50% node died

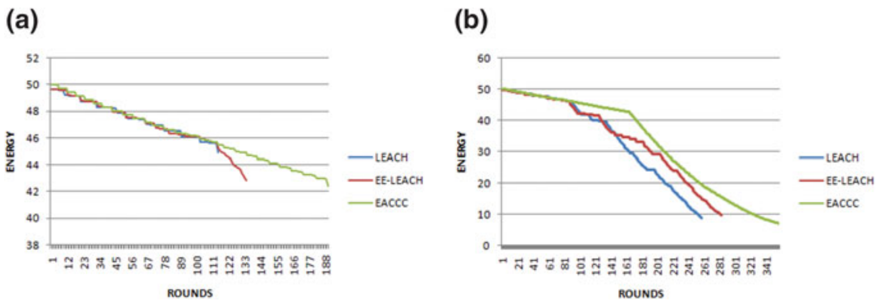


Fig. 5 Total network energy for homogeneous nodes. **a** First node died. **b** 50% node died

longer than that in LEACH, and EE-LEACH for the homogenous and heterogeneous nodes at first node died or 50% of nodes died.

Total Network Energy. Energy consumption can be calculated by the formula (3). Figures 5 and 6 show that the total network energy in EACCC is more than in LEACH, and EE-LEACH which means the energy consumption is minimized in EACCC in the case of the first node died or 50% nodes died for the homogeneous nodes or heterogeneous nodes.

Number of Failed Nodes. The node can be considered as the failed (dead) node, if the energy of sensor node is less or equal than 0. Figures 7 and 8 show that the failed

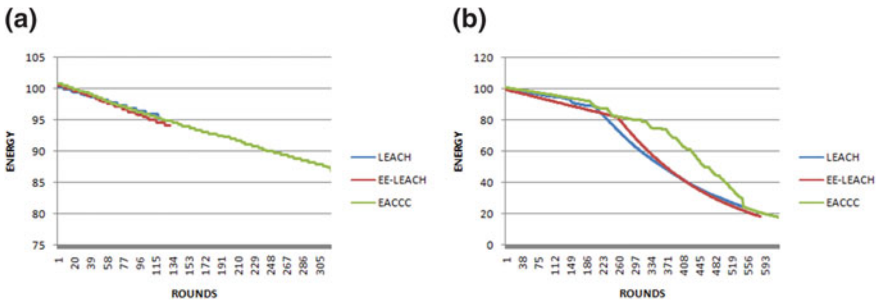


Fig. 6 Total Network energy for heterogeneous nodes. **a** First node died. **b** 50% node died

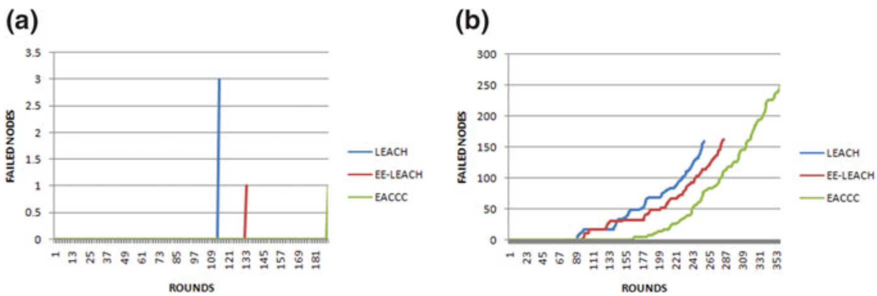


Fig. 7 Number of failed nodes for homogeneous nodes. **a** First node died. **b** 50% node died

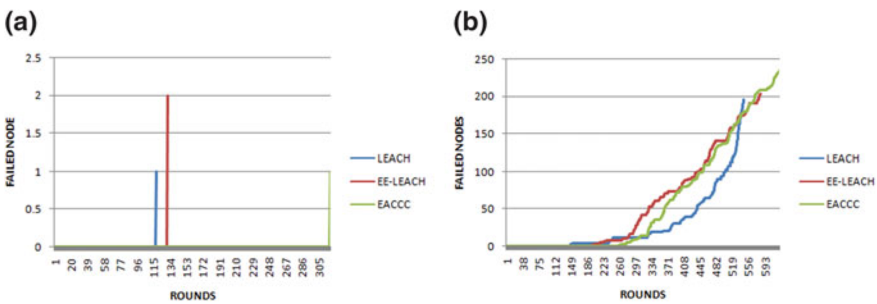


Fig. 8 Number of failed nodes for heterogeneous nodes. **a** First node died. **b** 50% node died

nodes number in the EACCC is less than in LEACH, and EE-LEACH in another word the alive nodes number is more in the EACCC in the case of first node died or 50% nodes died for the homogeneous nodes or heterogeneous nodes.

End-to-End Time Delay. It can be defined as the delay average between the data packet sending by the source and the same data packet receiving at the specific receiver, with the delay due to route acquisition, retransmission delays, and processing at intermediate nodes [17]. Figures 9 and 10 show the time delay between the LEACH, EE-LEACH and EACCC. The EACCC takes less time to aggregate

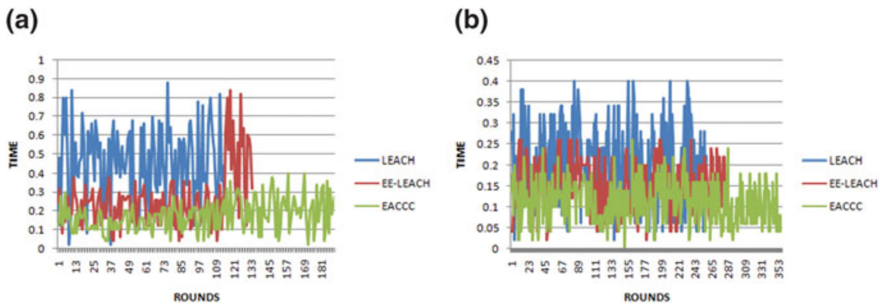


Fig. 9 End-to-end time for homogeneous nodes. **a** First node died. **b** 50% node died

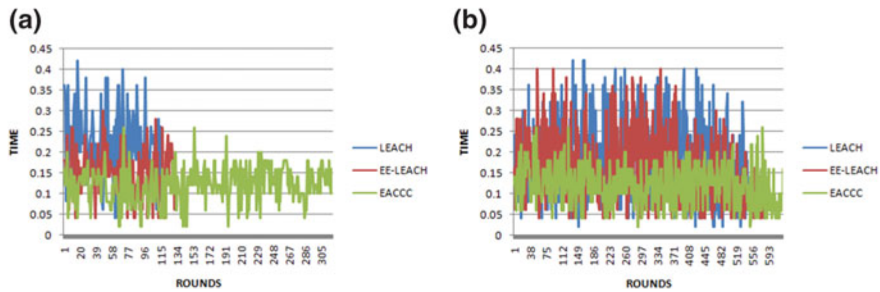


Fig. 10 End-to-end time for heterogeneous nodes. **a** First node died. **b** 50% node died

data and sends it to BS than LEACH, and EE-LEACH protocols for all cases of the simulation.

6 Conclusion

The main idea behind the design of protocol in WSNs is to keep the sensor nodes operating as possible, thus extending the lifetime of the network. In this paper, an improved method EACCC is proposed by using a centralized clustering algorithm for cluster heads selecting and the clusters forming in order to prolong the network lifetime and reduce the spent of the total energy in the network. The performance evaluation of EACCC is done through extensive analysis, analytical proof, and comparison. The results and performance analysis show that the EACCC has a better performance than LEACH protocol, and EE-LEACH protocol with respect to the lifetime of the network, total energy of network, No. of failed nodes, end-to-end delay, and routing overhead metric.

References

1. Al-Karaki, J.N., Kamal, A.E.: Routing techniques in wireless sensor networks: a survey. IEEE Wirel. Commun. **11**(6), 6–28 (2004)
2. Ilyas, M., Mahgoub, I.: Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems. CRC press (2004)
3. Nguyen, L.T., et al.: An energy efficient routing scheme for mobile wireless sensor networks. In: 2008 ISWCS'08. IEEE International Symposium on Wireless Communication Systems. IEEE (2008)
4. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: 2000 Proceedings of the 33rd annual Hawaii international conference on System Sciences. IEEE (2000)
5. Lindsey, S., Raghavendra, C.S.: PEGASIS: Power-efficient gathering in sensor information systems. In: 2002 IEEE Aerospace Conference Proceedings, vol. 3. IEEE (2002)
6. Arati, M., Agrawal, D.P.: TEEN: a routing protocol for enhanced efficiency in wireless sensor networks. In: IPDPS, vol. 1 (2001)
7. Heinzelman, W.B., Chandrakasan, A.P., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. IEEE Trans. Wirel. Commun. **1**(4), 660–670 (2002)
8. Bakaraniya, P., Mehta, S.: K-leach: An improved leach protocol for lifetime improvement in wsn. Int. J. Eng. Trends Technol. **4**(5), 1521–1526 (2013)
9. Arumugam, G.S., Ponnuchamy, T.: EE-LEACH: development of energy-efficient LEACH Protocol for data gathering in WSN. EURASIP J. Wirel. Commun. Netw. **2015**(1), 1–9 (2015)
10. Sindhwani, N., Vaid, R.: V LEACH: an energy efficient communication protocol for WSN. Mechanica Confab **2**(2), 79–84 (2013)
11. Alnawafa, E., Marghescu, I.: MHT: Multi-hop technique for the improvement of leach protocol. In: 2016 15th RoEduNet Conference: Networking in Education and Research. IEEE (2016)
12. Alnawafa, E., Marghescu, I.: IMHT: Improved MHT-LEACH protocol for wireless sensor networks. In: 2017 8th International Conference on Information and Communication Systems (ICICS). IEEE (2017)
13. Tudose, D., Gheorghe, L., Tapus. N.: Radio transceiver consumption modeling for multi-hop wireless sensor networks. UPB Sci. Bull. Ser. C Electron. Eng. **75**(1), 17–26 (2013)
14. Nikolidakis, S.A., et al.: Energy efficient routing in wireless sensor networks through balanced clustering. Algorithms **6**(1), 29–42 (2013)
15. Smaragdakis, G., Matta, I., Bestavros, A.: SEP: A stable election protocol for clustered heterogeneous wireless sensor networks. Boston University Computer Science Department (2004)
16. Legakis, H., Mehmet-Ali, M., Hayes, J.F.: Lifetime analysis for wireless sensor networks. In: IEEE GLOBECOM 2008 Global Telecommunications Conference (2008)
17. Gelet, G.: Performance Evaluation of Wireless Sensor Network Routing Protocols for Critical Condition Monitoring Application. M.A. Thesis. Addis Ababa University, Oct 2007

Security Challenges and Solutions for Wireless Body Area Networks

K. R. Siva Bharathi and R. Venkateswari

Abstract Wireless Body Area Networks (WBANs) are special purpose Wireless Sensor Networks, which is used to provide competent communication solutions for health care and medicinal applications. The rapid technological advancements in the field of sensors, MEMS, and the wireless communication enable the design and implementation of Wireless Body Area Networks. The most prominent application of WBANs is in healthcare but it also finds its applications in consumer electronics, sports safety, lifestyle, defense, and much more. WBANs are usually smaller networks when compared to WSNs but still, they are vulnerable to a massive number of security attacks. In this paper, we provide an overview of the Wireless Body Area Networks (WBANs), its applications, and security aspects. Various security threats and their countermeasures in WBANs are discussed based on the latest reviews and publications.

Keywords WBAN · Security · Attacks · Healthcare

1 Introduction

WBAN provides a remote mechanism to monitor and collect the data from a patient's body. Since remote access and very critical and vital data are involved, these WBAN networks require very high-level security and privacy at the time of storage and processing. Thus, the implemented WBAN infrastructure should possess various security features and guarantee security, privacy, integrity, and confidentiality of the data at all the times. The WBAN uses numerous sensors to collect biosignals from the human body and transmits it to a central device for processing and storage. These

K. R. Siva Bharathi
Department of ECE, Sri Krishna College of Engineering and Technology, Coimbatore, India
e-mail: sivabharathi.48@gmail.com

R. Venkateswari (✉)
Department of ECE, PSG College of Technology, Coimbatore, India
e-mail: rvenkateswari@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_29

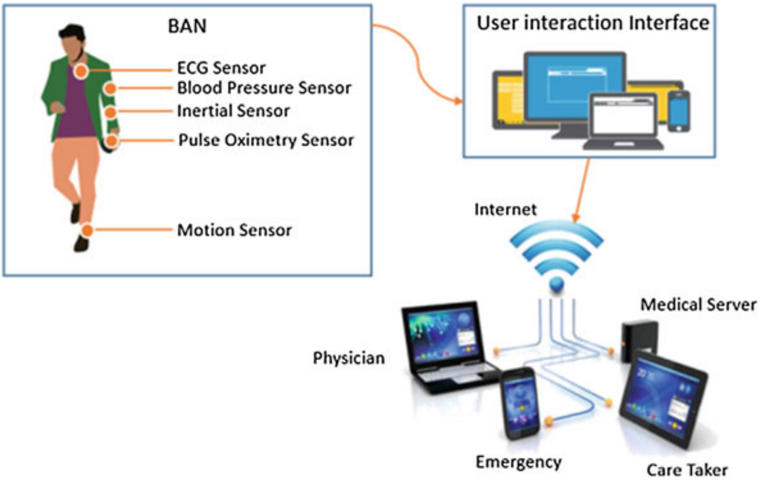


Fig. 1 Illustration of WBAN network

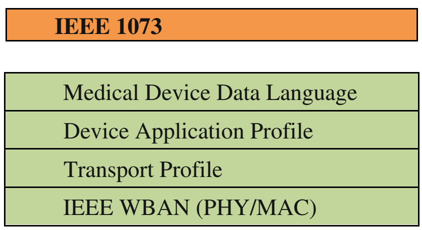
networks make use of the IEEE 802.15.6 standard and this standard classifies the WBAN nodes into three categories bestowing to the way they are implemented in the body and their role in the network [1]. A typical WBAN network is illustrated in Fig. 1.

The protocols and algorithms proposed for traditional Wireless Sensor Networks are not suitable for WBAN networks because of their unique features and applications. BAN networks are usually smaller networks, where the nodes are deployed in the human body. The node density is also much lower than compared to that of typical WSNs. Moreover, the data rate in BANs are more periodic and the BAN nodes are usually mobile [2].

A MobiHealth service platform enables remote monitoring of patients using wireless communication infrastructures. The IEEE 1073 [3] standard is modeled for Medical Device Communications and it acts as a new tool to health care providers. This standard provides a solution to wireless communication in WBAN. The IEEE 1073 model for WBAN is given in Fig. 2.

IEEE 802.15.6 is the latest standard defined for WBAN. The main purpose of IEEE 802.15.6 is to provide low power, short range, and reliable wireless communication

Fig. 2 IEEE 1073 model



supporting a vast range of data rates and various applications. This standard [4] defines three PHY layers, namely the Narrowband (NB), Ultra-Wideband (UWB), and Human Body Communications (HBCs). On the top of the physical layer, a refined MAC layer for access control to the channel is defined. Three levels of security are defined in this standard, namely unsecured communication, authentication but no encryption, authentication, and encryption.

The rest of the paper is organized as follows. The security aspects such as challenges and requirements are discussed in Sect. 2. Section 3 gives various security attacks and adopted solutions for the WBAN and Sect. 4 concludes the paper.

2 Security Aspects of WBAN Networks

WBAN faces serious and unique security threats when compared to other wireless networks. Even though a lot of security solutions are proposed, security in WBANs remains a greatest issue.

2.1 Security Challenges

The challenges arise in WBAN networks due to constrained resources such as power, energy, storage capacity, communication bandwidth, physical size, and nature of the data collected. The various challenges faced by the WBAN networks are discussed below.

Data Quality. The data obtained from a WBAN is very vital and life critic, and plays a important role in the health care process. All the further processes of the network depend on the data collected. Any inaccuracy in the data results in serious consequences. Hence, the quality of data collected must be of high quality and standard, and ensure the users to take the best decisions [5].

Data management. Various types of sensors are deployed in a WBAN for measuring different parameters. Also, a very huge amount of data is generated. Managing all these data and storing becomes the biggest challenge [6].

Data consistency. Data from multiple users from different locations have to be collected and processed. The end user of the data should contain all the necessary information regarding the collected data or else the final result might degrade [7].

Sensor validation. The sensor nodes in the network face serious communication issues and hardware constraints such as limited power and energy efficiency. Moreover, the sensor data is of more importance in WBANs. Hence, the sensor nodes have to be validated periodically [8].

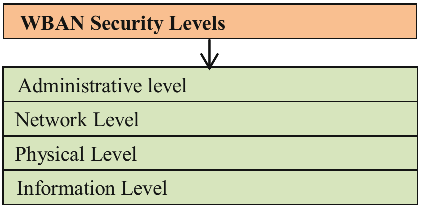


Fig. 3 Levels of security in WBAN

2.2 Security Requirements

Figure 3 illustrates the various levels of security [9] to be implemented in a WBAN system. Administrative level security includes the data access control, accountability, and revocability.

It also includes the authentication measures and audit of all entities. This enables to overcome the eavesdropping effect which ultimately results in node failure. The network level security includes maintaining the confidentiality of the data and non-repudiation of the data. The physical level of security includes resilience to node capture and Jamming attack. These are for inculcating security efforts to the attacks targeting the physical sensor devices. Information level security focuses on providing security to the obtained information. This is established through proper encryption, authentication, and freshness protection. The basic security requirements are understood by the CIA analysis, i.e., analyzing the confidentiality, integrity, and authentication of the data.

Data Confidentiality. It is the most important issue in WBANs. The data is to be protected from disclosure to any node other than the desired node. Many cryptographic algorithms are developed to ensure confidentiality of the data but the symmetric key cryptography is more reliable in WBANs as other algorithms are too costly to be implemented in resource constraint nodes.

Data Integrity. Apart from keeping the data confidential, the integrity of the data has to be maintained for a secure implementation. Loss of integrity occurs if a malicious node adds extra fragments or manipulates the data within a packet.

Data Authentication. Authentication is the procedure of confirming the original source of the data. IEEE 802.15.6 provides Authenticated Key Exchange (AKE) and Password-Based Authenticated Key Exchange (PAKE) protocols specific to the authentication [10].

3 Security Attacks and Adopted Solutions for WBAN

Figure 4 illustrates the various attacks encountered in different layers of WBANs. In each layer, a specific solution should be provided as explained below.

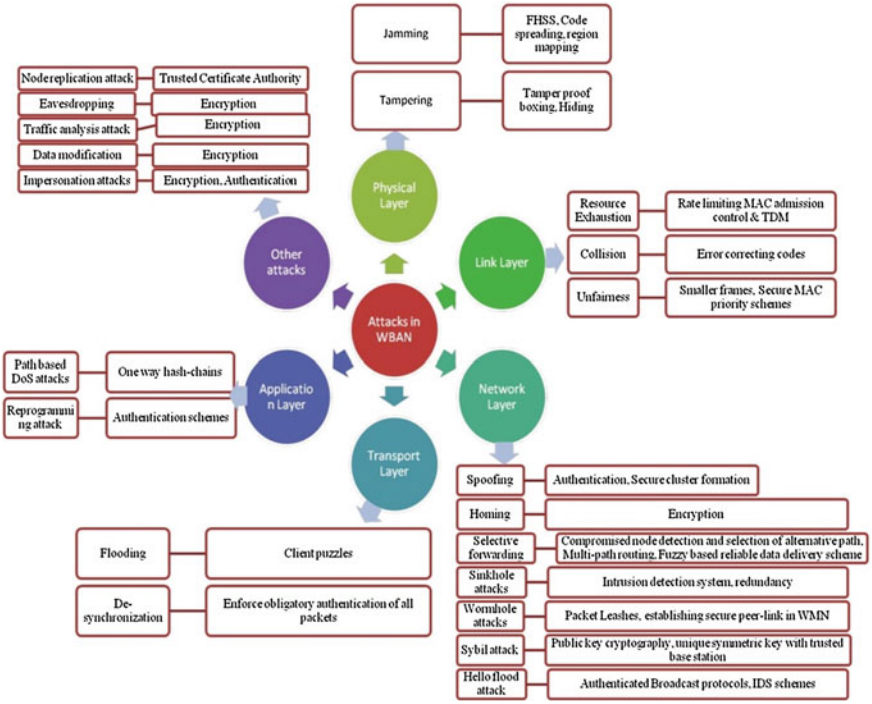


Fig. 4 Security attacks and solutions in the different layers of WBAN

3.1 Attacks in the Physical Layer

The most common attacks [11] in the physical layer are Jamming and Tampering. Jamming refers to the transmission of a random radio signal by the intruders and this radio signal interferes with the actual signal. Various solutions such as spread spectrum, Code spreading, Random Codekey Selection using Codebook (RCSC) DHSS [12], and honeynodes [13] are employed to overcome this. Tampering refers to exploitation of the cryptographic key and change the information or even replace the node with an adversary [14]. In WBANs, this is overcome by creating awareness to the patients regarding handling of the devices.

3.2 Attacks in the Link Layer

The data link layer is responsible for framing, multiplexing, error detection, and correction. The various attacks in this layer include resource exhaustion, collision, and unfairness in resource allocation.

3.3 Attacks in the Network Layer

The network layer is responsible for routing. Routing is done through the sink nodes for multiple WBAN networks. Possible attacks in this layer include spoofing, selective forwarding, sinkhole, wormhole, Sybil, and hello flood attacks. Selective forwarding is the attack in which an intruder drops packets in the network as it likes and then selectively forwards it to the neighbor nodes [15]. Techniques like Multipath routing, Multidata flow, and fuzzy-based reliable data delivery scheme are employed to overcome selective forwarding attacks [16]. Sinkhole is a type of attack where a malicious node attracts all the data packets in the network averting them from reaching the original destination [17]. Cumulative Acknowledgement-Based Detection (CADE) scheme is used to detect selective forwarding attacks but is used to prevent sinkhole attacks [18]. A wormhole attack is the one in which an adversary node keeps all the packets in one location and tunnels them to some other location [19]. Sybil attack is the one in which a single node pretends to have multiple identities in the network [20], and this is prevented by Message Authentication and Passing (MAP) algorithm. This algorithm makes use of the Compare and Match-position verification method for attack identification and applies MAP for prevention [21]. In hello flood attack, malicious node sends a HELLO packet to the network with a powerful radio transmission trying to convince all other nodes in the network to route packets through it [22]. One technique to overcome the HELLO flood attacks is the modified centralized IDS scheme based on misuse detection of the network resources [23].

3.4 Attacks in the Transport Layer

The transport layer is responsible for end-to-end delivery of the packets. The attacks encountered in the transport layer are flooding and desynchronization. A flooding attack is the one in which an attacker broadcasts many connection establishment requests to the victim node in the network thus depleting all the power of its resources. A desynchronization attack is the one in which the attackers desynchronize the endpoints of the network resulting in the nodes transmitting packets repeatedly for infinite times wasting their energy.

3.5 Attacks in the Application Layer

The application layer is responsible for providing services directly to the user applications. The application layer undergoes an especially, damaging form of the DoS attack, the Path-Based DoS (PDoS) attack. Here, an adversary node overwhelms the sensor nodes to a longer distance by flooding a multi-hop communication either

Table 1 Survey of security systems for WBAN networks

Security systems	Security metrics	Adopted Solutions
Internet of things for health care monitoring [25]	Confidentiality	Encryption using symmetric key cryptography
	Authentication	Node authentication model within the server
Trustworthy eHealth monitoring system [9]	confidentiality, no repudiation	Biometric cryptography for inter-WBAN communications
	encryption, freshness protection	Lightweight security mechanisms
	Authentication	Patient-centric authentication schemes
Pervasive social network-based healthcare [26]	Authentication, confidentiality	Authenticated association protocol
	Authentication integrity	Adding data to the block chains
Future internet systems design and implementation [27]	Trust, security, and privacy	Cloud-based IoT security module
Data Security and Privacy in WBAN [28]	Confidentiality	Attribute-based encryption (one-to-many encryption)
	Anonymity	Blind Signature
A systematic approach for IoT security [29]	Confidentiality and authentication	Automated key management scheme
Secure sensors data acquisition and communication protection in eHealthcare [30]	Authentication	AES-CCM hashing algorithm
	Authentication and encryption	AES encryption and CBC-MAC hashing algorithm
Secure mutual authentication for two-tier WBANs [31]	Mutual authentication, and anonymity	Anonymous authentication protocol and BAN logic

by replaying or injecting spurious packets [24]. Usage of one-way hash chains and proper authentication schemes prevent the various attacks in the application layer.

Table 1 lists the various security system developed for the wireless body area networks. It briefs the survey of various security metrics in WBANs such as Authentication, Confidentiality, Integrity, Trust, Privacy, and the adopted security solutions in them. The various security mechanisms implemented in WBANs are Cryptography, Key management, Secure routing, Resilience to node capture, Secure Localization, Trust management, and Robustness to communication denial of Services. The survey clearly explains that for a secure implementation of WBAN, all the security metrics have to be considered and to be dealt carefully for precise performance of the network.

Despite the deployment of WBAN, it still requires more research and development. Many wearable devices such as wristwatch, clothing jacket, and ring are miniatures. Fusing complex security protocols in this is a challenge. For implanted nodes, these issues are tougher [32]. As these devices are in close proximity to humans, interference and wireless RF hazards should be limited.

4 Conclusion

WBAN is a promising technology that has revolutionized the health care and lifestyle of people. Security plays a major role here as the network carries very sensitive information that has to be maintained confidentially. This paper surveys on the various security aspects, challenges, and threats a WBAN network is exposed to and the adopted countermeasures to almost all of the encountered attacks. With the development of Internet of Things (IoT), the WBANs will undergo a new revolution and its applications can be more interesting and exciting.

References

1. Samaher, A.J., Ibrahim, A.S., Shojafar, M., Shahaboddin, S.: Survey of main challenges (security and privacy) in wireless body area networks for healthcare applications. Egypt. Inform. J. **18**, 113–122 (2017)
2. Jamil, Y.K., Mehmet, R.Y.: New developments n Bio-medical Engineering, 2nd edn. InTech, Croatia (2010)
3. Kennelly, R.J.: The IEEE 1073 for medical device communications. In: IEEE Systems Readiness Technology Conference AUTOTESTCON '98. IEEE (2002)
4. Toorani, M.: On vulnerabilities of the security association in the IEEE 802.15.6 Standard. In: International Conference on Financial Cryptography and Data Security, vol. 8926, pp. 245–260 (2015)
5. Donoghue, J., Herbert, J., Sammon, D.: Patient sensors: a data quality perspective. In: International Conference on Smart Homes and Health Telematics, pp. 54–61. Springer (2008)
6. Donoghue, J., Herbert, J.: Data management within mHealth environments: patient sensors, mobile devices, and databases. J. Data Inf. Qual. **4**(1), 5:1–5:20 (2012)
7. Lai, D., Begg, R.K., Palaniswami, M.: Healthcare Sensor Networks: Challenges Towards Practical Implementation. CRC Press (2011)
8. Donoghue, J., Herbert, J., Fensli, R., Dineen, S.: Sensor validation within a pervasive medical environment. In: 5th IEEE Conference on Sensors, pp. 972–975 (2006)
9. Ajmal, S., Djahel, S., Zhang, Z.: Toward energy-efficient and trustworthy eHealth monitoring system. In: Wireless Communication over Zigbee for Automotive Inclination Measurement-China Communications (2015)
10. Toorani, M.: Security analysis of the IEEE 802.15.6 standard. Int. J. Commun Syst **29**, 2471–2489 (2016)
11. Pejman, N.: Wireless body area networks: attacks and countermeasures. Int. J. Sci. Eng. Res. **6**(9) (2015)
12. Balamurugan, G., Bhagyaveni, M.A.: Random Codekey Selection Using Codebook without Pre-Shared Keys for Anti-jamming in WBAN, vol. 51. Elsevier (2016)

13. Mishra, S., Dhurandher, S., Rayankula, A., Agrawal, D.: Using honeynodes for defense against jamming attacks in wireless infrastructure-based networks. Comput. Electr. Eng. **36**(2), 367–382 (2010)
14. Saleem, S., Ullah, S., Yoo, H.S.: On the security issues in wireless body area networks. Int. J. Digit. Content Technol. Appl. **3**(3) (2009)
15. Deng, H., Sun, X., Wang, B., Cao, Y.: Selective forwarding attack detection using watermark in WSNs. In: International Colloquium on Computing, Communication, Control and Management. IEEE (2009)
16. Hae, Y.L., Tae, H.C.: Fuzzy-based Reliable Data Delivery for Countering Selective Forwarding in Sensor Networks. Springer, pp. 535–544(2007)
17. Tumrong, W.C., Varakulsiripunth, R.: Detecting sinkhole attack and selective forwarding attack in wireless sensor networks. In: International Conference on Information Communication and Signal Processing. IEEE (2009)
18. Yu, B., Xiao, B.: Detecting selective forwarding attacks in wireless sensor networks, In: Proceedings of the 20th International Parallel and Distributed Processing Symposium (2006)
19. Hu, Y.C., Perrig, A., Johnson, D.B.: A defence against wormhole attacks in wireless networks. In: Proceedings of 22nd Annual Joint Conference of the IEEE Computer and Communications, vol. 3 (2003)
20. Newsome, J., Shi, E., Song, D., Perrig, A.: The sybilattack in sensor networks: analysis & defenses. In: Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks (2004)
21. UdayaSuriya, R.D., Rajamani, V.: Detecting and preventing sybil attacks in wireless sensor networks using message authentication and passing method. Sci. World J. **2015** (2015)
22. Hamid, M.A., Mamun, O., Rashid, M., Hong, C.S.: Routing security in sensor network: Hello flood attack and defense. In: IEEE ICNEWS, pp. 2–4 (2006)
23. Kaur, P., Gurm, J.S.: Detect and prevent HELLO FLOOD attack using centralized technique in WSN. Int. J. Comput. Sci. Eng. Technol. **7** (2016)
24. Deng, J., Han, R., Meshra, S.: Defending against Path-based DoS attacks in wireless sensor networks. In: Proceedings of the 2005 ACM Workshop on Security of Ad Hoc and Sensor Networks (2005)
25. Anass, R., Fatiha, E., Bouhorma, M.: The internet of things for healthcare monitoring: security review and proposed solution. In: Third International IEEE Colloquium on Information Science and Technology (2015)
26. Zhang, J., Xue, N.: A Secure System for Pervasive Social NetworkbasedHealthcare, Article in IEEE Access (2016)
27. Sotiriadis, S., Petrakis, E.: Future Internet Systems Design and Implementation: Cloud and IoT Services Based on IoT-A and FIWARE. Springer (2017)
28. Li, M., Lou, W.: Data security and privacy in wireless body area networks. IEEE Wirel Commun. 1536–1284 (2010)
29. Riahi, A., Challal, Y., Natalizio, E., Chtourou, Z., Bouabdallah, A.M.: A systemic approach for IoT security. In: IEEE International Conference on Distributed Computing in Sensor Systems, (2013)
30. Yaseen, M., Saleem, K., Derhab, A., Abbas, H., Muhtadif, J.A.: Secure sensors data acquisition and communication protection ineHealthcare: review on the state of the art, Telemat. Inform. (2017)
31. Ibrahim, M.H., Kumari, S., Das, A.K., Wazid, M., Odelu, V.: Secure anonymous mutual authentication forstar two-tier wireless body area networks. J. Comput. Methods Progr. Bio Med. **135**, 37–50 (2016)
32. Zou, S., Xu, Y., Wang, H., Li, Z., Chen, S., Bo, H.: A Survey on Secure Wireless Body Area Networks Hindawi Security and Communication Networks (2017)

Minimizing Congestion in Mobile Ad hoc Network Using Adaptive Control Packet Frequency and Data Rate

Navneet Kaur and Rakesh Singhai

Abstract An effective congestion control algorithm should ensure reliable message delivery, quality of service, and energy optimization. This paper presents a method in which data rate and control packets such as Hello packet interval is selected according to the channel conditions depending on the node mobility and energy consumed by the nodes in transmission. The method enables nodes to adjust their data rate and frequency of their Hello messages depending on the transmission power and current speed of the nodes. The improved protocol detects and reacts to congested parts of the network by using adaptive data rate and Hello packet interval. This helps in reducing congestion and improve throughput. The functionality of the proposed method is tested using the Network simulation tool. The results have been analyzed in various scenarios to evaluate the performance parameters. This mainly improves throughput and reduces end-to-end delay and jitter in high mobility cases. The average queue length is also controlled.

Keywords Congestion control · Hello interval · MANET · Traffic control
Fuzzy logic · Adaptive data rate

1 Introduction

Dynamic changes in network topology are experienced in ad hoc networks because of free movement of the nodes. Links between two nodes may break during the data transmission due to movement of intermediate nodes [1]. The link can also get disrupted due to scarce node energy. More is the transmission power of the node, more is the energy consumed by the node in transmitting packet. Both these factors causes failure of the node and results in the re-establishment of the route. This results in broadcasting of control packets for route rediscovery, which increases the traffic

N. Kaur (✉) · R. Singhai
Department of Electronics and Communication, UIT RGPV, Bhopal, India
e-mail: navkaur2005@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_30

and causes congestion. Congestion in a network results when the total demand for a resource such as link bandwidth, exceeds the capacity of the resource [2].

There are two basic approaches to solve the problem of controlling congestion in wireless network, namely: traffic control [3–9] and resource control [10–13]. The application requirements are met by employing either traffic control to throttle the node rates [4] or resource control by exploiting idle resources [13]. The traffic control adapts transmission rate to available resources capacities in order to avoid congestion. Source nodes reduce the traffic load they generate, to respond to congestion build up. A rate based congestion control scheme is proposed in [5] which is based on rate based generic Additive Increase Multiplicative Decrease and Additive Increase Additive Decrease scheme, depending on the congestion status. The limitation of this approach is that congestion is notified by using special packets that are communicated through common control channel with high priority. This consumes network resources and thus the method has long latency.

These traffic control algorithms result in less power consumption of a node and minimal packet drops. The major weakness of this scheme is that the bandwidth is underutilized in case of low traffic, whereas when the traffic load is high, the delay is large [8, 9]. None of the above papers [4–7] on traffic control technique are able to immediately regulate the traffic according to channel conditions.

Reducing source traffic may affect application requirements, thus resource provisioning techniques can be used as an alternative to rate control methods. In order to handle high resulted traffic, it is better to increase the capacity by turning on more resources [10]. The results show that resource control or alternative path creation algorithms ensure high lifetime of a network but these load balancing techniques causes extra overhead for congestion notification.

The congestion control protocols suggested in the above papers [4–13] overlook the problem of additional traffic emitted in the network. The above shortcomings are addressed in this paper which proposes a fuzzy logic based adaptive method and prolongs the network lifetime by reducing the amount of traffic sent in the network. The objective of this paper is to solve the problem of traffic control by adapting the data rate and also reduces the load on the network due to periodic broadcasting of Hello packets.

This paper is organized as follows. The overview is presented in Sect. 2. The problem identification and proposed methodology is discussed in Sect. 3. The performance analysis and results is detailed in Sect. 4. Finally, Sect. 5 concludes this work.

2 Overview

In Ad hoc On-demand Distance Vector (AODV) routing protocol, nodes periodically broadcast Hello packets to proclaim their presence and location to their neighbors. When a node receives a Hello message, there exists a viable communication channel between a node and the source of the Hello message, thus Hello message is used to determine nodes connectivity in a network. If a beacon is not received by a node from one of its neighbours within a certain time interval, the corresponding node is assumed to have left the transmission range due to its mobility [14].

The energy consumed by the node in transmission is proportional to transmission power of a node. The energy consumed in transmitting a packet (Etx) [15] can be computed as

$$Etx = \frac{Ptx * Psize}{BW}, \tag{1}$$

where Psize represents the packet size, Ptx is the packet transmitting power and bandwidth of the link is represented by BW. As the energy of the node gets depleted, the link will be disrupted. This causes packet drops and further enhances the network congestion. Data sending rate can be reduced to mitigate congestion, but the bandwidth is underutilized when the traffic load is minimal. Whereas when the traffic load is high, the delay is large. The data rate can be made adaptive on indications of congestion to avoid this situation.

3 Proposed Work

3.1 Problem Identification

Each node transmits Hello packets periodically to announce its presence to its neighbours. However, the adaptive Hello packet updates are more significant than employing updation by using static beaconing packets. Several drawbacks of periodic beacon packets are discussed as following:

- Periodic packet beaconing consumes energy of the node as the beacon packets are transmitted, received, and processed by nodes. Hello packets can collide with user data packets which results in loss of data.
- Mobile Ad hoc networks have a limited bandwidth. User data packets will be dropped if the bandwidth is used by Hello packets, thus effectively reducing overall performance of a network.

If Hello interval is high, there is delay in the detection of link failure event and the neighborhood tables become obsolete. On the contrary, if it is too low, then tables will be updated but the congestion in the traffic will be high. There should exist a tradeoff between routing efficiency and extra overhead which in turn depends on network characteristics that are dynamic. Thus a constant Hello interval is not a best choice.

Many studies have evaluated performance of Ad hoc routing protocols based on the utilization of Hello messages. Researchers have focused on the correlation between topology changes due to node mobility and Hello interval. To the best of my knowledge, all existing solutions use the fuzzy rules that are based on the

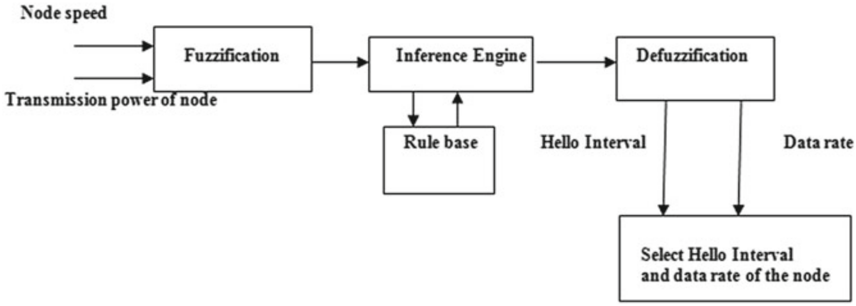


Fig. 1 Block diagram for the basic components of proposed fuzzy system

assumption that Hello interval is made adaptive according to the changeable topology due to nodes movements. In previous researches, Hello interval varies inversely with node speed. High speed of a node results in a disruption of link with some of its current neighbors. Therefore, the HI time of those nodes is made small to send more Hello messages to record the links breaks. These approaches overlook the problem of congestion due to broadcasting of control packets. But the approach taken for framing rules in this paper is different. One of the key aspects of our work is that HI is adaptively optimized according to node speed in order to avoid congestion.

3.2 *Proposed Methodology*

A fuzzy controlled Data rate based and Hello interval based congestion control is proposed in this research. The rules for fuzzy Hello interval and data rate that is used with AODV are introduced. The fuzzy rules consider node mobility and transmission power as input variables and Hello interval and data rate as output variables in two different cases as shown in the block diagram given in Fig. 1. As per the output of fuzzy rules, after detecting and differentiating the congestion level, system is used for selecting the optimal Hello interval and data rate according to the detected congestion status.

The description of fuzzy rules for adaptive Hello interval and data rate is explained in following sections.

3.2.1 Effect of Node Speed and Transmission Power on Data Rate

Assume the node has a high speed; it results in retransmission of control and data packets due to frequent breakage of link. High mobility of node will cause network congestion in the node neighborhood, thus node brings down its rate of transmitting data.

Table 1 Rule table for calculation of data rate

Transmission power → Speed ↓	Low (L)	Medium (M)	High (H)
Low (L)	High	Medium	Low
Medium (M)	High	Medium	Low
High (H)	Medium	Medium	Low

- In case of low speed, there is less change in the network topology and route is maintained. Thus the proposed rule is that data rate is kept high in case of less mobility of nodes and it is reduced for high mobile environment.
- When the transmission power of a node is high, the energy consumed by the node is more, resulting in expiry of node due to energy depletion. The congestion will be more due to link failure and thus data sending rate should be decreased to reduce congestion. In case of node transmitting with low transmission power, the energy consumed is less and node will have longer lifetime. The data rate can be increased as stable route is maintained. Thus the rule follows that data rate is low for high transmission power and it is high for low transmission power.
- Similarly, the other rules are framed. According to the first rule, if node speed and transmission power is low, there will be no congestion and it is desirable for a node to transmit at high data rate.
- According to the rule, if node speed is high and transmission power is low, and then data rate is medium. The reason is that more routes will become invalid due to high mobility. Whereas the energy consumption of nodes is low as data is transmitted with low power, thus link breakage due to energy elimination is reduced. Thus the congestion is not a very severe problem and data rate is selected medium.
- All the nine rules can be combined to adapt data rate based on node speed and transmission power. The rule table to control data rate adaptively is given in Table 1.

3.2.2 Effect of Node Speed and Transmission Power on Hello Interval

- In addition to controlling the Data rate according to node speed, we improve our mechanism by adapting the Hello interval to the node mobility. Since the link lifetime of slow moving node with its neighbors is large. There are chances of reduced packet drops and retransmission of packets, which results in low congestion. In this condition, Hello packets are frequently broadcasted by the node without affecting the congestion.
- In contrast, the highly unpredictable topology, due to nodes movement, should decrease the frequency of Hello messages because mobility can cause link breakage and increased traffic. The broadcast of Hello packets get wasted when a node does not participate in communication, thus Hello packets should not be unnecessarily broadcasted for connectivity information. Thus it follows the rule, that

Table 2 Rule table for calculation of Hello interval

Transmission power → Speed ↓	Low (L)	Medium (M)	High (H)
Low (L)	Low	Medium	High
Medium (M)	Low	Medium	High
High (H)	Medium	High	High

when the speed of node is low, the Hello interval will be low whereas it will be high for high speed.

- When the transmission power of a node is too low, the energy consumed by a node is less and the possibility of route failure due to energy elimination of that node reduces. The congestion due to packet loss and retransmission of packets is less. Thus, the Hello interval is small enough as it will not affect the network efficiency. When transmission power of the node is low, Hello interval is low and it is high in case of high transmission power. The rest of the rules are formulated according to combined effect of node speed and transmission power.
- When the speed of the node is high, more routes become invalid and new control packets are propagating in network for route discovery. Packets are dropped as buffers get full and, thus links are about to be congested. At the same time, the low transmission power of the node results in less consumption of node energy. The Hello interval is kept medium as it will take care of channel conditions. Thus follows the rule, when speed is high and transmission power is low, the Hello interval is medium.

The rules for calculation of Hello Interval are given in Table 2.

The proposed fuzzy rules are summarized in Tables 1 and 2, which are used by fuzzy inference to map speed and transmission power input sets into data rate and Hello interval output sets.

4 Results and Analysis

The data rate and Hello message is made adaptive according to the channel conditions The selected values of node speed and node transmission power and the corresponding values of Hello interval and data rate obtained through fuzzy logic system are implemented in Qualnet simulator. The results for various performance parameters such as throughput, end-to-end delay, jitter and average queue length are plotted and compared.

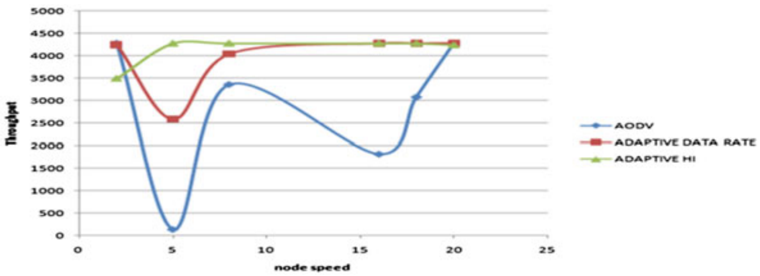


Fig. 2 Graph of throughput versus Node speed

4.1 Simulation Configuration

Simulations are conducted on the QualNet simulator 5.0. Multiple simulations are run for simulation scenario to obtain results, which show improved performance on all observed metrics. The simulation parameters are summarized as follows:

- In the scenario, 20 nodes are randomly placed in an area of 1500 * 1500 m. Each node moves towards a randomly chosen destination node with a random speed. The maximum node speed in the scenario is 20 m/s and the minimum node speed is 2 m/s and the pause time is 30 s. The total length of the simulation is 300 s.
- In network, nine sources are used to generate data traffic and size of the data packet is 512 bytes.
- The transmission range is 300 m, and MAC layer protocol used is 802.11 DCF protocol. The node mobility is expressed by the maximum speed. Membership functions for different parameters used in fuzzy rules consider different ranges as per the settings in Qualnet simulator.

4.2 Results

The performance comparison analysis of the AODV and proposed variants is discussed in this section. The performance of adaptive algorithm has been analyzed for change in the speed of the nodes. In this work, the data rate is made adaptive according to congestion status so that packets transmitted in the network will be controlled. Thus, rate of packet drop reduces and throughput is improved as compared to basic AODV as illustrated in Fig. 2. Similarly, the frequency of Hello packets is made adaptive so that traffic due to broadcasting of control packets is controlled according to channel conditions. Thus, the nodes have longer lifetime and are able to forward packets for a longer period of time, due to energy saving. Thus, throughput is more as the number of received packets is more for the proposed algorithm as shown in Fig. 2.

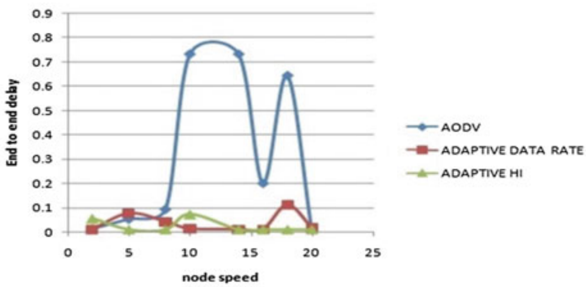


Fig. 3 Graph of end-to-end delay versus Node speed

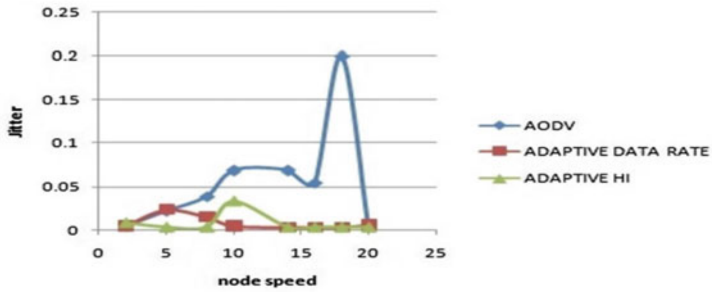


Fig. 4 Graph of jitter versus Node speed

Figures 3 and 4 demonstrate the end-to-end delay and jitter, respectively. The graph shows that the end-to-end delay and jitter are high in AODV due to frequent route failure. Congestion increases with the increase in control and data traffic and packet have to wait in the queue and would take more time to reach to destination. The proposed adaptive technique reduces route disconnection rate and propagation of control packets, which reduces possibility of congestion. The Hello overhead is significantly reduced and improves delay and jitter. The average queue length of the nodes is also reduced in the proposed variants due to reduced packet drops and traffic as shown in Fig. 5. This is due to adaptive transmission of control packets.

From Figs. 2, 3, 4 and 5 graphs, it is apparent that delay, jitter and average queue length is reduced in case where Hello interval and data rate is made adaptive in comparison to traditional AODV.

5 Conclusion

A fuzzy controlled rate based and Hello interval based congestion control is proposed to enhance the performance of mobile network in terms of parameters throughput, end-to-end delay, jitter and average queue length. The sending data rate is declined

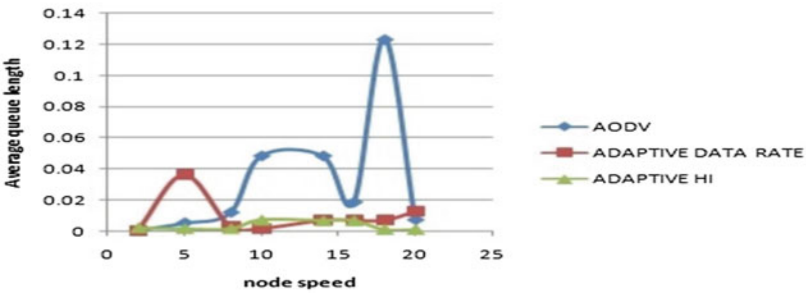


Fig. 5 Graph of average queue length versus Node speed

upon occurrence of congestion, so that congestion is avoided and is not experienced by intermediate nodes. This ensures a reliable communication within Mobile networks. Hello message is not broadcasted periodically and its frequency is decided according to mobility and transmission power of the nodes.

Simulation results demonstrate that all the performance parameters are improved for both the proposed variants but highest improvement is shown in the case where Hello messages are made adaptive. The performance parameters outperformed in case of adaptive Hello interval as compared to adaptive data rate. Thus, it is concluded that congestion can be controlled more efficiently by modifying and controlling the frequency of Hello messages.

References

1. Johnson, D.B., Maltz, D.A.: Dynamic source routing in ad hoc wireless networks. Mobile Computing, vol. 10, no. 7, pp 153–181. Springer, USA (1996)
2. Jain, R.: Congestion control in computer networks: Issues and trends. IEEE Netw. **4**(3), 24–30 (1990)
3. Yang, P., Shao, J., Luo, W., Xu, L., Deogun, J., Lu, Y.: TCP congestion avoidance algorithm identification. IEEE/ACM Trans. Netw. (TON) **22**(4), 1311–1324 (2014)
4. Paek, J., Govindan, R.: RCRT: rate controlled reliable transport protocol for wireless sensor networks. ACM Trans. Sens. Netw. **7**(3), pp. 20:1–20:45 (2010)
5. Esmaeelzadeh, V., Hosseini, E.S., Berangi, R., Akan, O.B.: Modeling of rate-based congestion control schemes in cognitive radio sensor networks. Ad Hoc Netw. Elsevier **36**, 177–188 (2016)
6. Pu, J., Hamdi, M.: Enhancements on router-assisted congestion control for wireless networks. IEEE Trans. Wirel. Commun. **7**(6), 2253–2260 (2008)
7. Kliazovich, D., Granelli, F.: Cross-layer congestion control in ad hoc wireless networks. Ad Hoc Netw. Elsevier **4**(6), 687–708 (2006)
8. Bansal, G., Kenney, J.B.: LIMERIC: a linear adaptive message rate algorithm for DSRC congestion control. IEEE Trans. Veh Technol. **62**(9), pp. 1–1 (2013)
9. Feng, C.W., Huang, L.F., Xu, C., Chang, Y.C.: Congestion control scheme performance analysis based on nonlinear RED. IEEE Syst. J. (2015)
10. Kafi, M.A., Djenouri, D., Ben-Othman, J., Badache, N.: Congestion control protocols in wireless sensor networks: a survey. IEEE Commun. Surv. Tutor. **16**(3), 1369–1390 (2014)

11. Tran, D.A., Raghavendra, H.: Routing with congestion awareness and adaptivity in mobile adhoc networks. IEEE Trans. Parallel Distrib. Syst. **17**(11) (2006)
12. Mallapur, S.V., Siddarama, R.P., Jayashree, V.A.: Load balancing technique for congestion control multipath routing protocol in MANETs. Wireless Personal Communications, vol. 92, Issue 2, pp 749–770. Springer (2017)
13. Senthil kumaran, T., Sankaranarayanan, V.: Dynamic congestion detection and control routing in manet. J. Comput. Inf. Sci. Elsevier **25**(1), 25–34 (2013)
14. Heissenbuttel, M., Torsten, B.: Optimizing neighbor table accuracy of position-based routing algorithms. In: IEEE INFOCOM (2005)
15. De Rango, F., Guerriero, F., Fazio, P.: Link-stability and energy aware routing protocol in distributed wireless networks. IEEE Trans. Parallel Distrib. Syst. **23**(4), 713–726 (2012)

Network Selection Scheme Using Taguchi Method for Real-Time Streaming Media Over Heterogeneous Networks

Renuka Deshpande, Lata Ragha and Satyendra Kumar Sharma

Abstract Next generation wireless communication networks needs to integrate various heterogeneous technologies based on an IP core network. Thus it guarantees service continuity, optimum network selection, user mobility and integration of new applications and resources. Integration of various wireless technologies in heterogeneous environment offers best service to every application. However, an automatic interface selection and user preferences based on quality of service parameters such as available resources, bandwidth, network delay or speed, security, and power consumption is desired. Hence network selection scheme explicitly based on user preferences and available resources is required. In this paper, network selection scheme using Taguchi method over heterogeneous wireless communication networks is proposed. Taguchi method assists in analyzing the quality of service parameters so as to satisfy or establish best optimum parameters of the network that further assist in network selection. It also estimates the contribution from the individual parameters affecting the overall performance of the network. Moreover, our results ensure optimum selection of suitable matching network for every flow considering quality of service parameters and user preferences. Network selection scheme finds applications in the area of real-time streaming media, online traffic offloading, reduced delay, optimum utilization of the services, integration of heterogeneous wireless technologies, incorporation of new applications and quality of services through user preferences.

Keywords Network selection scheme · Taguchi method · Quality of service parameters · User preferences

---

R. Deshpande (✉)
Faculty of Engineering, Pacific Academy of Higher Education and Research University, Udaipur, India
e-mail: renukagdeshpande@gmail.com

L. Ragha
Department of Computer Engineering, Fr. C. Rodrigues Institute of Technology, Vashi, Navi Mumbai, India

S. K. Sharma
Modern Institute of Technology, Research Centre, Alwar, India

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_31

1 Introduction

Tremendous growth in wireless networks such as WiFi, WiMax, LTE, etc., has enabled many applications taking advantage of mobility. Therefore it has become extremely difficult to satisfy all application needs such as available resources, bandwidth, network delay, security and power consumption. Next generation wireless communication networks needs to integrate various heterogeneous technologies based on an IP core network. Thus it guarantees service continuity, optimum network selection, user mobility, and integration of new applications and resources. Integration of various wireless technologies in heterogeneous environment offers best service to every application. Practically several factors affect the network selection scheme. Hence network selection scheme explicitly based on user preferences and available resources is required. Heterogeneous network, mobility management issues, network characteristics, etc., makes network selection scheme extremely difficult. To overcome above problem many researchers have proposed various approaches based on mobile terminal, access network or both. Mechanism proposed at network side to assist mobile terminal on network selection based on compensatory and non-compensatory multi-attribute decision [1–3]. Also, many mechanisms that optimize power consumption and user satisfaction have been suggested but found to be very complex [4–7]. Recently, many researchers have proposed network selection schemes that are based on user preferences, application and network characteristics [8–14]. However, these schemes are found to be unsatisfactory since they do not consider the mobile terminal characteristics. Therefore, it is almost necessary to consider mobile terminal decision without operator intervention that satisfies its requirements. However, to ensure satisfactory quality of service to all mobile terminals, it is mandatory to share network characteristics that may be sometimes dangerous from security point of view. But other network characteristics such as available bandwidth, speed, capacity, operator policy and network conditions are shared to the mobile terminals. Moreover, to accommodate mobile terminal's characteristics such as battery status, memory and CPU may unnecessarily complicate the network selection scheme. Thus it is required to proposed network selection scheme based on user preferences, application requirements and network characteristics. Mobile terminal characteristics need to be overlooked considering the limited computing resources available at the terminal.

In this paper, network selection scheme using Taguchi method over heterogeneous wireless communication networks is proposed. Taguchi method assists in analyzing the quality of service parameters so as to satisfy or establish best optimum parameters of the network that further assist in network selection. It also estimates the contribution from the individual parameters affecting the overall performance of the network. Moreover, our results ensure optimum selection of suitable matching network for every flow considering quality of service parameters and user preferences. Network selection scheme finds applications in the area of real-time streaming media, online traffic offloading, reduced delay, optimum utilization of the services, integration of heterogeneous wireless technologies, incorporation of new applications and quality of services through user preferences. The paper is organized as follows Sect. 1 introduces to network selection scheme, Sect. 2 describes the proposed algorithm for network selection based on user preferences through Taguchi method in details. Results are discussed in Sect. 3 and finally concluded in Sect. 4.

2 Network Selection Scheme Based on Taguchi Method

Wireless networks are gradually adopting heterogeneity from its current homogeneity in order to support many applications, quality of service parameters and user preferences. Mobile terminals are equipped with capabilities to access available networks, various protocol stacks, applications and quality of service parameters. The next generation network infrastructure is supposed to provide IP based multi-service network such as WiFi, WiMax, LTE and evolving technology. In this method, it is assumed that the mobile terminal is responsible for network selection based on the user preferences and network characteristics available through data collection and analysis. System architecture of the mobile terminal for heterogeneous network environment remains similar and that network selection process has been proposed. Other logical and functional modules such as security, call administration, messaging are not discussed and beyond the scope of this paper.

2.1 Network Selection Policies

Network selection scheme in heterogeneous environment is challenging due to its dependency on many factors such as bandwidth, speed, signal strength, power consumption, user satisfaction and interference. Every network selection schemes have their own advantages and disadvantages. Mostly these network selection schemes are designed to satisfy certain user/mobile terminal requirements in terms of bandwidth, speed, and power consumption. Network selection schemes do not interfere with system performance and are not involved in call administration, security, etc. In this algorithm, we consider the access networks available at mobile terminals and needs access while satisfying the following requirements.

1. Maximum bandwidth offered by the network
2. Minimum power consumption at the mobile terminal
3. Better security levels
4. Minimum delay
5. Accept user preferences in terms of specific requirement based on bandwidth, security levels, and speed.

Network selection scheme takes the required network characteristics as its input and selects the best network satisfying above criteria. The scheme selects the best among the available networks and does not involve in enforcement of the result. It is also necessary to block inputs to the network selection scheme periodically so as avoid falsely triggered to the algorithm. Thus it selects the network only when required and does not increase power consumption, reduce battery life and takes CPU time of mobile terminal.

2.2 System Model

In this section, we determine the model for selection of the network for a particular application based on Taguchi's method. Taguchi Method is primarily a R&D development method for experimentally improving quality. It is achieved by performing experiments in which the various parameters are varied and noise is added to determine the quality of the product or method. Thus it generates best control parameters to achieve desired results. In this experiment for network selection scheme, network parameters such as bandwidth, speed, power consumption, and security levels are included for all the available networks. Variations in these parameters from minimum to maximum are included as noise. User preferences are included as desired quality or parameters in the experiment. Thus it is expected that the result of this method would select the best possible network that closely matches with user preference parameters. A quality of service (QoS) request needs to be generated by the mobile terminal from all the available networks before access network. The QoS contains bandwidth, speed, security levels and power consumption requirements for the network. Let A be the application for which network access is required with following user preferences in terms of QoS parameters.

B_d Desired bandwidth of network for application A
V_d Desired delay of network for application A
S_d Desired security levels for application A
P_d Desired power consumption at mobile terminal for application A

Let $N = \{N_1, N_2, N_3, N_4\}$ are the set of heterogeneous networks available at the mobile terminal to execute the desired application A with following QoS parameters.

BN_i^{min} *Minimum bandwidth available for network*
BN_i^{max} *Maximum bandwidth available for network i*
VN_i^{min} *Minimum delay available for network i*

Table 1 Orthogonal array L9

Experiments	QoS parameters			
	BN	VN	SN	PN
1	BN^{min}	VN^{min}	SN^{min}	PN^{min}
2	BN^{min}	VN^{mid}	SN^{mid}	PN^{mid}
3	BN^{min}	VN^{max}	SN^{max}	PN^{max}
4	BN^{mid}	VN^{min}	SN^{mid}	PN^{max}
5	BN^{mid}	VN^{mid}	SN^{max}	PN^{min}
6	BN^{mid}	VN^{max}	SN^{min}	PN^{mid}
7	BN^{max}	VN^{min}	SN^{max}	PN^{mid}
8	BN^{max}	VN^{mid}	SN^{min}	PN^{max}
9	BN^{max}	VN^{max}	SN^{mid}	PN^{min}

VN_i^{max} *Maximum delay available for network i*
SN_i^{min} *Minimum security levels available for network i*
SN_i^{max} *Maximum security levels available for network i*
PN_i^{min} *Minimum power consumption at mobile terminal for network i*
PN_i^{max} *Maximum power consumption at mobile terminal for network i*

The one variation in those QoS parameters between minimum and maximum is considered due to noise. Thus, each QoS parameters for four networks have three levels of variations that would be considered in this experiment. These QoS parameters with their variations for each network will be analyzed to satisfy user preference parameters through Taguchi's method. In this experiment considering four QoS parameters with each having three levels thus it takes exactly 64 experiments to determine the suitability of the network for application A. Taguchi's method helps to reduce the number of experiments to nine using orthogonal array L9 [15]. Table 1 shows the orthogonal array L9 for design of experiments. Thus Taguchi's method helps in accelerating the process of analyses through reducing the number of experiments. At the same time, it guarantees desired results without any or minimum error.

Thus to determine the network throughput at various QoS parameters for four networks it would logically take 256 experiments. Also, the number of experiments will linearly increase with addition of each network. Thus Taguchi's method reduces the number of experiments to nine that results in 36 experiments to determine the

Table 2 Orthogonal array L9 for network 1 with throughput

Experiments	QoS parameters				Network throughput (Mbps)	S-P
	BN	VN	SN	PN		Product
1	10	10	3	10	52.43	30
2	10	45	2	30	11.65	60
3	10	100	1	50	5.24	50
4	45	10	2	50	52.43	100
5	45	45	1	10	11.65	10
6	45	100	3	30	5.24	90
7	100	10	1	30	52.43	30
8	100	45	3	50	11.65	150
9	100	100	2	10	5.24	20

throughput at all possible QoS parameters and four heterogeneous networks. Thus this scheme reduces computation that inherently reduces CPU time, power consumption and memory requirement at mobile terminal during execution of network selection scheme.

3 Experimental Results

The performance of the network selection scheme based on Taguchi's method was measured by calculating the network throughput. The network throughput was calculated using bandwidth-delay product for TCP buffer size of 64 KB. In this experiment to incorporate security levels and power dissipation in the network throughput, we also calculated security-power product. The security-power product needs to be minimum since expected power dissipation is minimum and security level one is considered as higher and three as lower. Orthogonal array L9 was derived for all the networks with their QoS parameters and throughput. Table 2 shows the orthogonal array L9 for network 1 along its throughput derived from bandwidth-delay (BW-D) and security-power (S-P) product. Table 2 consists of QoS parameters as per their values obtained through network characteristics. It is populated using Taguchi's orthogonal array L9 that is used for design of experiments or measurement of results. Similarly, the desired throughput for QoS user preference parameters was calculated and tabulated in Table 3. Orthogonal array L9 was obtained for all the other networks 2, 3 and 4 which is not shown but can be calculated using Table 1. Taguchi's method is not limited to only four parameters and three levels, but instead can be scaled up as required. Therefore this method can add any number of QoS parameters and its level for reducing the number of experiments.

Table 3 Desired throughput for user preference QoS parameters

QoS parameters				Desired throughput (Mbps)	Throughput
BN	VN	SN	PN		S-P
100	10	1	30	52.43	30

Table 4 QoS parameters for all the networks

Network 1	Network 2	Network 3	Network 4
$BN_1^{min} = 10$ MHz	$BN_1^{min} = 1$ MHz	$BN_1^{min} = 10$ MHz	$BN_1^{min} = 1$ MHz
$BN_1^{mid} = 45$ MHz	$BN_1^{mid} = 5$MHz	$BN_1^{mid} = 45$MHz	$BN_1^{mid} = 5$MHz
$BN_1^{max} = 100$MHz	$BN_1^{max} = 10$MHz	$BN_1^{max} = 100$MHz	$BN_1^{max} = 10$MHz
$VN_1^{min} = 10$ms	$VN_1^{min} = 10$ms	$VN_1^{min} = 100$ms	$VN_1^{min} = 5$ms
$VN_1^{mid} = 45$ms	$VN_1^{mid} = 45$ms	$VN_1^{mid} = 150$ms	$VN_1^{mid} = 12$ms
$VN_1^{max} = 100$ms	$VN_1^{max} = 100$ms	$VN_1^{max} = 200$ms	$VN_1^{max} = 25$ms
$SN_1^{min} = 3$	$SN_1^{min} = 3$	$SN_1^{min} = 3$	$SN_1^{min} = 3$
$SN_1^{mid} = 2$	$SN_1^{mid} = 2$	$SN_1^{mid} = 2$	$SN_1^{mid} = 2$
$SN_1^{max} = 1$	$SN_1^{max} = 1$	$SN_1^{max} = 1$	$SN_1^{max} = 1$
$PN_1^{min} = 10$mW	$PN_1^{min} = 30$mW	$PN_1^{min} = 40$mW	$PN_1^{min} = 90$mW
$PN_1^{mid} = 30$mW	$PN_1^{mid} = 60$mW	$PN_1^{mid} = 60$mW	$PN_1^{mid} = 120$mW
$PN_1^{max} = 50$mW	$PN_1^{max} = 90$mW	$PN_1^{max} = 80$mW	$PN_1^{max} = 150$mW

Desired throughput for user preference QoS parameters and network bandwidth for all the networks can be compared for data transfer (Table 4). The minimum differences in throughput and available bandwidth would finally result in best available network for application A. After determining network throughput for all the networks and desired throughput, it is clearly understood that to satisfy QoS user preferences parameters network 1 is more suitable since it has minimum delay of 10 ms and maximum bandwidth of 100 MHz. Therefore it results in satisfactory network throughput and faster data transfer as compared with other networks. Above method can be effectively applied for real-time streaming media applications in heterogeneous networks that require large amount of data transfer. It offers flexibility in selecting appropriate network for user application. Also, selection of network can be periodically reviewed for performance and updated or changed if required.

4 Conclusion

In this paper, we proposed network selection scheme based on Taguchi's method in heterogeneous network environment. An automatic interface selection and user preferences based on quality of service parameters such as available resources, bandwidth, network delay or speed, security, and power consumption are desired. Taguchi's method provides simple, computationally less complex and optimized solution for network selection scheme. The scheme may result in low power consumption and faster network selection. The proposed network selection scheme is suitable due to increase in the number of applications and real-time streaming media over heterogeneous networks.

References

1. Qadeer, M.A., Ahmad, R., Khan, M.S., Ahmad, T.: Real-time video streaming over heterogeneous networks. In: International Conference on Advanced Communication Technology, vol. 2, Mar 2009
2. Pudlewski, S., Cen, N., Guan, Z., Melodia, T.: Video transmission over lossy wireless networks: a cross-layer perspective. IEEE J. Sel. Top. Signal Process. **9**(1), 6–21 (2015)
3. Sen, J., Bhattacharya, S.: A survey on cross-layer design frameworks for multimedia applications over wireless networks. Int. J. Comput. Sci. Inf. Technol. **1**(1), 29–42 (2008)
4. Luo, H., Shyu, M.-L.: Quality of service provision in mobile multimedia—a survey. Human-centric Comput. Inf. Sci. A Springer Open J. **1**(1) (2011)
5. Chen, J.-Y., Chiu, C.-W., Li, G.-L., Chen, M.-J.: Burst-aware dynamic rate control for H.264/AVC video streaming. IEEE Trans. Broadcast. **57**(1), 89–93 (2011)
6. Egilmez, H.E., Civanlar, S., Tekalp, A.M.: An optimization frame work for QoS-enabled adaptive video streaming over openflow networks. IEEE Trans. Multimed. **15**(3), 710–715 (2013)
7. Balk, A., Maggiorini, D., Gerla, M., Sanadidi, M.Y.: Adaptive MPEG-4 video streaming with bandwidth estimation. In: Quality of Service in Multiservice IP Net works, Second International Workshop, QoS-IP 2003, Feb 2003
8. Ji, W., Li, Z., Chen, Y.: Joint source-channel coding and optimization for layered video broadcasting to heterogeneous devices. IEEE Trans. Multimed. **14**(2), 443–455 (2012)
9. Dreibholz, T., Becke, M., Pulinthanath, J., Rathgeb, E.P.: Implementation and evaluation of concurrent multipath transfer for SCTP in the INET framework. In: 3rd International ICST Conference on Simulation Tools and Techniques, Mar 2010
10. Liao, J., Wang, J., Li, T., Zhu, X.: Introducing multipath selection for concurrent multipath transfer in the future Internet. Elsevier Comput. Netw. **55**, 1024–1035 (2011)
11. Jacob, J.S., Preetha, K.G.: A QoS based network selection scheme for 4G systems with preplanner for emergency services. In: IEEE International Conference on Computing, Communication and Networking Technologies, India (2013)
12. Kuboniwa, J., et al.: High efficient network selection scheme using location information for heterogeneous wireless system. In: IEEE International Wireless Distributed Networks Workshop on Cooperative and Heterogeneous Cellular Networks, USA (2015)
13. Awad, A., Mohamed, A., Chiasserini, C.-F.: Dynamic network selection in heterogeneous wireless networks. IEEE Consum. Electron. **6**(1), 53–60 (2017)

14. Hwang, R.H., Cheng, K.C.: Novel QoS guaranteed cell selection schemes in LTE-A heterogeneous networks. In: 2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media), Pattaya, pp. 1–5 (2017)
15. Roy, R.: A primer on the Taguchi's method. In: Van Nostrand Reinfold International Company Limited, 1990

Performance of Internal Cluster Validations Measures For Evolutionary Clustering

Pranav Nerurkar, Aruna Pavate, Mansi Shah and Samuel Jacob

Abstract Clustering is an NP-hard grouping problem and thus there are advantages of using a metaheuristic (swarm intelligence) strategy to find the near global optimal solution to it. To effectively guide the agents of the swarm in the metaheuristic strategy, a suitable cost function is needed for successful outcome. The current inquiry focuses on the use of internal validation criteria as cost functions as they achieve the dual goals of clustering which are compactness and separation. Out of the multiple internal validation criteria included in the literature, two are identified for this purpose, viz. BetaCV and Dunn index. These were used as cost functions of the swarm optimizer metaheuristic (PSO-BCV and PSO-Dunn). To demonstrate the validity of the proposed technique, it was compared with other metaheuristics differential evolution as well as the traditional swarm optimizer based on distance-based criteria (PSO). The analysis of the results obtained on clustering benchmark datasets highlighted the suitability of this approach.

The original version of this chapter was revised: The affiliation of authors Mansi Shah and Samuel Jacob has been updated along with the city name in the affiliation of author Aruna Pavate. The correction to this chapter is available at https://doi.org/10.1007/978-981-13-1513-8_105

P. Nerurkar (✉)
Department of CE & IT, VJTI, Mumbai, India
e-mail: pranavn91@gmail.com

A. Pavate
Department of CE & IT, Atharva CoE, Mumbai, India
e-mail: gavkare@gmail.com

M. Shah
Department of CE & IT, Rizvi CoE, Mumbai, India
e-mail: mansishah928@gmail.com

S. Jacob
Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu, Rajasthan, India
e-mail: samueljacob@atharvacoe.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_32

Keywords Evolutionary clustering · Swarm intelligence · Cluster analysis
Cluster validation · Optimization

1 Introduction

Cluster validation measures (CVI) are used to understand the presence or absence of clustering tendency and arrive at a "good" clustering solution [1]. These measures are useful as the techniques which are applied on the data might not be able to find natural groupings present in the data [2]. CVI are of three types, namely internal cluster validation measures, external cluster validation measures, and relative validation measures. The first type does not require clustering information to be provided with the data. It relies on statistics calculated from the clustering solution obtained by the unsupervised algorithm. It specifies a range bound value (cost) of the obtained clustering solution. This value is an indicator of whether the results of the clustering are suitable. Thus, such measures can be used to decide whether to draw inferences from the data or not [1, 2]. External validation indices are useful, if ground truth labels are provided with the data, these find application in both classification (supervised learning) as well as clustering (unsupervised learning) [3]. Purity, Information gain, Rand index, and Variation of information are examples of this type. Relative clustering measures are more suitable for the task of finding the optimal number of clusters in the data. As most of the real-world data will not have class labels associated with it, internal validations indices could be more useful in assessment of clustering results.

Clustering algorithms reviewed in the literature belong to partitioning based algorithms [4] as well as hierarchical clustering techniques. Both these categories of algorithms use heuristics to assign data points to their respective prototypes or centroids [5]. Clustering the data is an NP hard problem and no heuristic based approach can traverse through the solution space to arrive at the global maxima. Hence, this investigation argues for the suitability of evolutionary algorithms for the task of identifying the optimal clustering of data [6].

Evolutionary algorithms traverse the search space in a parallel manner. The inherent structure of such algorithms lead to avoidance of local optima in the search space [6]. The second advantage of evolutionary algorithms is that they allow consideration of multiple solutions to the same data which is not the case seen in nonevolutionary algorithms [7]. However, the success of a suitable metaheuristic strategy depends on the selection of a suitable cost function. The role of the cost function is to evaluate the solutions generated by the metaheuristic at an iteration. This helps in determining which swarm particle is at the best position in that iteration. This information can be used by the metaheuristic to align the individual swarm agents towards the current best position and determine their next course.

Use of internal validation criteria as a cost function has dual advantages. First, it removes the need to test the validity of clustering results obtained by the algorithm. Second, as internal cluster validation indices rely on maximizing both separation and compactness, they might be able to avoid biases seen in single objective optimization strategies. In this paper, it is argued that a swarm optimizer based on validation criteria would perform better than one on traditional distance based heuristics. The reasoning behind this intuition is that use of such indices would convert the problem of clustering into a dual-objective optimization problem. This would lead to a better clustering, otherwise, a technique based on a single objective would be biased towards detection of only a single type of cluster. The above approach could not be found in the literature review.

2 Related Work

2.1 Internal Validation Criteria

Calinski-Harabasz index [1, 8] It is the average inter- and intra-cluster sum of squared distances. A high value indicates compact and well-separated clusters.

$$\frac{\sum_i n_i d^2(c_i, c)/(NC - 1)}{\sum_i \sum_{x \in c_i} d^2(x, c_i)/(n - NC)} \tag{1}$$

I Index [1, 8] It is the ratio of maximum distance between centroids and sum of distances of objects to the centroid of their cluster. Ideally, it should have a maximum value.

$$(\frac{1}{NC} * \frac{\sum_{x \in D} d(x, c))}{\sum_i \sum_{x \in c_i} d(x, c_i)} * max_{i,j} d(c_i, c_j))^P \tag{2}$$

Dunn's Indices [1, 8] Ratio of minimum intercluster distance to the maximum intra-cluster distance. Ideally, it should be high.

$$min_i(min_j(\frac{min_{x \in c_i, y \in c_j} d(x, y)}{max_k(max_{x,y \in c_k} d(x, y))})) \tag{3}$$

BetaCV [1, 8] Ratio of mean intra-cluster distance to mean intercluster distance. A smaller value indicates better clustering.

$$\frac{d_{intra}}{d_{inter}} \tag{4}$$

$$d_{intra} = avg_{x \in c_i, y \in c_i} d(x, y)) \tag{5}$$

$$d_{intra} = avg_{x \in c_i, y \in c_j} d(x, y)) \tag{6}$$

Silhouette index [1, 8] Higher value is ideal as it indicates how similar a value placed in a cluster is to other members of its own cluster when compared to members belonging to other clusters.

$$\frac{1}{NC}\sum_{i}\left(\frac{1}{n_i}\sum_{x\in C_i}\frac{b(x)-a(x)}{max[b(x)-a(x))]}\right) \tag{7}$$

Davies Bouldin index [1, 8] For every point, the similarity value to each cluster is calculated. The highest of these values is given to the point. The DB-index of the data is the average of the values of all points. Smaller value indicates clusters are distinct from each other.
Xie Beni index [1, 8] It is ratio of minimum square distance between centroids and mean square distance between each point and its centroid. Lower value is ideal.

$$[\sum_{i}\sum_{x\in C_i} d^2(x, c_i)]/[n * min_{i,j\neq i} d^2(c_i, c_j)] \tag{8}$$

where D : Dataset, C_i: ith cluster, n_i: number of points in C_i, c_i: centroid of C_i, P: number of attributes in D, NC: number of clusters, $d(x, y)$: L_2 norm distance between x and y, $a(x)$: average dissimilarity of x with all other data in same cluster, $b(x)$: lowest average dissimilarity of x to other clusters to which it does not belong to.

Other validation indices such as symmetry distance-based index $Sym - index$ is applicable to only datasets that are internally symmetric, whereas composite density between and within-cluster index $CDbw$ gives unstable results on some datasets as it may not find centroids for each cluster [8]. For these reasons, these are excluded from this section.

3 Mathematical Model

Dunn's validation index is conceptually the simplest of the internal validation indices and suitable to be a cost function [9]. BetaCV was also a suitable candidate for a cost function as it considers mean intra-cluster and intercluster distances and shall be less sensitive to outliers in data compared to indices like Xie-Beni and I-index as they rely on minimum or maximum distance based criteria [8].

3.1 Cost Function—BetaCV and Dunn index

Algorithm 1: Compute Dunn Index

```
Result: Returns Dunn index of the data
distMat = euclidean distance matrix of data;
ind = column vector of cluster ids;
minInter = 10000;
maxIntra = -1;
for i ← 1 to max − clust − 1 do
    for j ← (i + 1) to max − clust do
        temp = minWeights(distMat, ind, i, j);
        if temp < minInter then
            minInter = temp;
for i ← 1 to max − clust do
    temp = maxWeights(distMat, ind, i, i);
    if temp > maxIntra then
        maxIntra = temp;
result = minInter / maxIntra;
```

Algorithm 2: Compute BetaCV Index

```
Result: Returns BetaCV index of the data
distMat = pdist2(X, X);
ind = column vector of cluster ids;
Win = 0;
for i ← 1 to max − clust do
    Win = Win + sumWeights(distMat, ind, i, i);
Wout = 0;
for i ← 1 to max − clust − 1 do
    for j ← (i + 1) to max − clust do
        Wout = Wout + sumWeights(distMat, ind, i, j);
Nin = 0;
for i ← 1 to max − clust do
    n = sum(ind == i);
    Nin = Nin + (n * (n-1) / 2);
Nout = 0;
for i ← 1 to max − clust − 1 do
    for j ← (i + 1) to max − clust do
        Nout = Nout + sum(ind == i) * sum(ind == j);
result = (Win / Nin) / (Wout / Nout);
```

Table 1 Description of the datasets

Sr. No.	Name	Size	Attributes	Classes
1	E. coli	336	7	8
2	Wisc	628	9	2
3	Wine	178	13	3
4	Iris	150	4	3
5	Yeast	1484	8	9

The above algorithms are used as cost functions to guide the swarm- based meta-heuristic as described in Sect. 2.

4 Experimental Study

Real datasets from fields such as Cellular Biology [EColi], Cancer detection [Wisc], Plant species [Iris], Wine types [Wine], and yeast families [Yeast] are used for the cluster analysis. Swarm optimizer using Dunn index (PSO-Dunn) and BetaCV (PSO-BCV) as cost functions are compared with the Particle Swarm Optimizer [PSO] that uses distance-based heuristics [10–12] as well as other metaheuristic such as differential evolution [13]. Distance-based statistics obtained from clusters such as widest within-cluster gap, average silhouette width, and external validity criteria, viz. Variation of Information index [14] and Corrected Rand index [15] are used as metrics to measure performance.

4.1 Dataset

The datasets used for the experiment are given in Table 1.

4.2 Results

Distance-based Statistics Widest within-cluster gap (Wgap): BetaCV and Dunn index-based clustering has higher *Wgap* on three of the five datasets compared to other metaheuristics. As both measures improve compactness, *Wgap* might increase as close points are clustered together (Table 2).

Silhouette value (SH): BetaCV and Dunn index-based clustering has higher *SH* on four of the five datasets compared to other metaheuristics. This could be as both measures seek to improve compactness (Table 3).

Table 2 Widest within-cluster gap

Name	E. coli	Wisc	Wine	Iris	Yeast
PSO	0.55	9.16	133.22	0.81	0.56
PSO-Dunn	**0.72**	**9.16**	133.2	0.73	0.58
PSO-BCV	0.54	**9.16**	133.2	**0.82**	0.5
DE	0.55	9.16	133.2	0.82	0.6

Table 3 Average silhouette width

Name	E. coli	Wisc	Wine	Iris	Yeast
PSO	0.27	0.6	0.57	0.55	0.21
PSO-Dunn	0.23	0.53	**0.6**	0.51	0.34
PSO-BCV	**0.29**	0.6	0.57	**0.55**	**0.48**
DE	0.25	0.6	0.57	0.53	0.28

Table 4 Variation of information

Name	E. coli	Wisc	Wine	Iris	Yeast
PSO	1.11	0.31	1.23	0.51	2.72
PSO-Dunn	**1.1**	0.67	1.06	0.51	2.37
PSO-BCV	1.26	0.33	1.24	0.55	**2.35**
DE	1.56	0.31	1.26	0.46	2.38

Table 5 Corrected rand index

Name	E. coli	Wisc	Wine	Iris	Yeast
PSO	0.56	0.86	0.39	0.74	0.07
PSO-Dunn	**0.62**	0.45	0.28	0.56	0
PSO-BCV	0.58	0.84	0.37	0.73	0
DE	0.51	0.86	0.37	0.8	0.13

External Validation Criteria : Variation of information (VI): BetaCV and Dunn index-based clustering has better *VI* on two of the five datasets compared to other metaheuristics (Table 4).

Corrected Rand Index (CRI): BetaCV and Dunn index-based clustering has higher *CRI* on one of the five datasets (Table 5).

5 Conclusion

Internal cluster validation indices have advantages as cost functions as they seek to achieve dual objectives of separation and compactness. This means clustering can become a dual objective optimization. However, this also introduces an additional overhead to the task of clustering due to the higher computational cost when compared to traditional distance-based heuristics used in clustering. Metaheuristic approach is used for clustering and due to the advantage, it provides over traditional clustering algorithms in terms of lower computation cost, parallel execution, faster convergence, effective traversal of search space, and the ability to avoid local optimal solutions. This paper utilizes two internal clustering validation criteria, viz. BetaCV and Dunn index for guiding the agents of the metaheuristic to reach the global optima. Both criteria are chosen due to their conceptual simplicity and robustness against outliers compared to other indices. The results have shown that the strategy provided in this paper can achieve effective clustering and results comparable in performance with other metaheuristics.

References

1. Liu, Y., Li, Z., Xiong, H., Gao, X., Junjie, W., Sen, W.: Understanding and enhancement of internal clustering validation measures. IEEE Trans. Cybern. **43**(3), 982–994 (2013)
2. Baya, A.E., Granitto, P.M.: How many clusters: a validation index for arbitrary-shaped clusters. IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) **10**(2), 401–414 (2013)
3. Guo, G., Chen, L., Ye, Y., Jiang, Q.: Cluster validation method for determining the number of clusters in categorical sequences. IEEE Trans. Neural Netw. Learn. Syst. (2016)
4. Rui, X., Wunsch, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**(3), 645–678 (2005)
5. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recogn. Lett. **31**(8), 651–666 (2010)
6. Chakrabarti, D., Kumar, R., Tomkins, A.: Evolutionary clustering. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 554–560. ACM (2006)
7. Hruschka, E.R., Campello, R.J.G.B., Freitas, A.A., et al.: A survey of evolutionary algorithms for clustering. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.), **39**(2), 133–155 (2009)
8. Han, J., Pei, J., Kamber, M.: Data Mining: Concepts and Techniques. Elsevier (2011)
9. Dalton, L., Ballarin, V., Brun, M.: Clustering algorithms: on learning, validation, performance, and applications to genomics. Curr. Genomics **10**(6), 430–445 (2009)
10. Eberhart, R., Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, 1995. MHS'95, pp. 39–43. IEEE (1995)
11. Bratton, D., Kennedy, J.: Defining a standard for particle swarm optimization. In: Swarm Intelligence Symposium, 2007. SIS 2007, pp 120–127. IEEE (2007)
12. Clerc, M.: Particle Swarm Optimization, vol. 93. Wiley (2010)
13. Storn, R., Price, K.: Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. **11**(4), 341–359 (1997)
14. Meilă, M.: Comparing clusteringsan information based distance. J. Multivar. Anal. **98**(5), 873–895 (2007)
15. Hennig, C.: How many bee species? a case study in determining the number of clusters. In: Data Analysis, Machine Learning and Knowledge Discovery, pp. 41–49. Springer (2014)

Performance Analysis of Polar Coded IHDAF Relaying for Next Generation Cellular Networks

N. Madhusudhanan and R. Venkateswari

Abstract Cooperative relay networks play a vital role in improving the coverage and capacity of cellular networks. In order to achieve high reliability, relaying should be used with channel codes. In this paper, the hybrid relaying protocol based on the incremental procedure is proposed through polar codes. It provides 40% gain when the threshold signal-to-noise ratio between source and relay equals 5 and threshold signal-to-noise ratio between source and destination is optimal. Simulation result illustrates that incremental HDAF using polar codes outperform in Alamouti scheme compared to single-input-single-output systems.

Keywords Incremental HDAF protocol · Bit error rate · Alamouti scheme

1 Introduction

1.1 Overview

Over the past decade, International Mobile Telecommunication Union (IMU) has developed cellular standards for global mobile communications. The mobile cellular industry today is patiently looking to upgrade the recent release of current cellular standards in terms of quality of service. Most important challenges faced by the standard are providing better throughput and high data rate at the cell edge. New communication technologies like multiple transmission and reception systems, multicarrier modulation schemes, error control coding techniques, relaying protocols and carrier aggregation has been considered to improve the performance in terms of capacity and coverage. In this situation, cooperative relaying plays an important

N. Madhusudhanan (✉) · R. Venkateswari
Department of ECE, PSG College of Technology, Coimbatore, India
e-mail: madhusudhanan.n83@gmail.com

R. Venkateswari
e-mail: rvenkateswari@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_33

role. Relaying is basically a low powered base station, transfers information from source to the destination [1–3]. With respect to 3GPP specifications, relaying protocols are basically classified into three types namely, Amplify-and-Forward (AF) relay, decode-and-forward (DF) relay, and compress-and-forward (CF) relay. In an amplify-and-forward relay, the desired signal will get amplified along with noise and forwards them to the destination. Decode-and-Forward (DF) relay decodes the desired signal and transfers the received signal to the destination with the elimination of noise present in AF through digital signal processing methods. In Compress-and-Forward (CF) relaying protocol, the incoming signal will get compressed and forwarded to the destination with a reduction in processing delay [2]. Other forms of DF are Fixed DF (FDF) and Adaptive DF (ADF). The working principle of FDF is similar to normal decode-and-forward relay. In the case of an adaptive decode-and-forward relay (ADF), the relay forwards the received signal to the destination when the system meets the threshold level. Then, hybrid cooperative relaying schemes have been introduced in cooperative communication systems. They are classified as Decode-Amplify-and Forward (DAF) relaying, Incremental Hybrid relaying protocol. This hybrid scheme exploits selective-cooperation to accept or reject the cooperation of relay. After hybrid relaying technique, the concept of incremental technique has been introduced to improve spectral efficiency by means of feedback mechanism. Positive acknowledgment will be sent to the source, if information reaches the destination otherwise negative acknowledgment will be sent to retransmit the information [4]. Hybrid Decode-Amplify-Forward relaying scheme has been investigated in terms of symbol error rate and performance gain with other conventional relaying schemes [5]. In [6], incremental relaying technique was proposed to achieve the performance in terms of error rate and outage behavior. The concept of an incremental decode-amplify-and-forward relay with selective cooperation over Rayleigh fading channels has been presented to estimate the performance of error rate [7]. Later, an incremental hybrid decode-amplify-and-forward relaying protocol was proposed in [8] to estimate outage probability and bit error rate with respect to signal-to-noise ratio threshold along with power constraints.

In cooperative communications, channel coding has been chosen as a candidate for error correction in many applications to achieve capacity [9]. During 2008, Erdal Arikan has introduced polar codes as a family of channel codes for different channels like binary discrete memoryless channel (B-DMC), additive white Gaussian noise channel (AWGN), Rayleigh and other fading channels [10]. In this coding scheme, successive cancelation (SC) decoder is utilized at the decoding part with reduced latency and complexity. Now, the polar code has been considered a major candidate for 5G cellular networks due to the features of providing high throughput and data rate with reduced latency [11]. The block error rate for polar coded spatial modulation (PCSM) over Rayleigh flat fading and Rician channels was analyzed in [12].

The main contribution of the proposed system is to reduce bit error rate for cooperative communication systems. The incremental HDAF (IHDAF) cooperative protocol is incorporated with polar codes to provide better error rate performance for single and multiple transmission and reception systems. In this paper, some of the capacity constraints like signal-to-noise ratio (SNR), bit error rate, threshold SNR are taken into consideration over Rayleigh channel.

The rest of the paper is organized as follows: Sect. 2 describes the proposed polar coded IHDAF (PCIHDAF) relaying system. Simulation results of the proposed system are shown in Sect. 3. Section 4 concludes the paper.

2 Proposed Polar Coded IHDAF Relaying System

Erdal Arikan invented a new channel code "Polar Codes" and these codes are highly capacity-achieving and have the error-correcting capability in which it approaches the Shannon limit in a closed manner with coding gain without any error floor [9, 11]. Channel polarization along with Bhattacharya parameter $Z(\mathrm{W})$ [1] has been utilized for large code length and elaborated by Matthew Effect. Then, the channels are divided into the following categories: noisy and noiseless channels [1]. The basic strategy is to transmit information over the channel without noise.

During the encoding operation, the set of frozen bits U_a is selected at the source. The corresponding frozen set of indices is given by [2]

$$U_a = \{i \in 0, 1, \ldots, N-1\} : Z\left(W_a^i\right) \leq \delta_N \tag{1}$$

where $Z\left(W_a^i\right)$ denotes the Bhattacharya parameter and signifies the process of construction for different channels and it is mainly dependent on combine and split operations. δ_N is the capacity bound parameter. On the other hand, the set of information bits are also present. The set of information indices is denoted by [2]

$$U_b = \{i \in 0, 1, \ldots, N-1\} : Z\left(W_c^{\mathrm{i}}\right) \leq \delta_N, \tag{2}$$

where $Z\left(W_c^{\mathrm{i}}\right)$ represents the Bhattacharya parameter of the ith channel between source and destination.

In the proposed system, the half-duplex cooperative relaying technique with IHDAF relaying system through polar codes is considered. It includes a source (S), a relay (R) and a destination (D) as shown in Fig. 1. It demonstrates the proposed system model for lower to higher order antenna systems and utilizes two antennas at the transmitter (NT_1, NT_2) and a single antenna at the receiver (NR_1). The channel coefficients such as h_a, h_b, and h_c are considered as channel gain between source-relay, source-destination and relay-destination, links respectively. All the channel gains of the system model obey Rayleigh density function.

At the initial stage, the information bits are passed through the system, which is encoded through polar codes and broadcasted by the source (S). The relay and destination will receive the desired information sent by the source. Basically, relay forwards the transmitted signal, only if the channel quality between source and relay is good and 'X' denotes the vector of bits applied to the Rayleigh channel by the

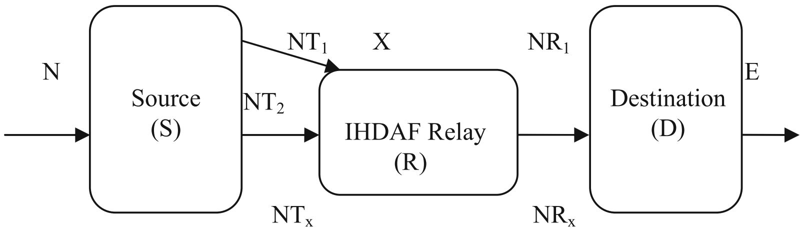


Fig. 1 System model of polar code based IHDAF relaying

relay. During the next stage, the source and relay cooperatively transmit signals to the destination through Alamouti scheme. In the proposed system, the polar encoder is used at the transmitter side and successive cancelation decoder at the receiving side. 'E' is the estimate to recover original information at the destination. This system will work either in DF or AF to estimate the maximum rate that can be maintained across the link for all possible channel states. In polar coded IHDAF relaying technique, broadcasting of information is carried out in two different stages. The source directly transmits data to the destination at the initial stage. The relay terminal kept silent when an instantaneous signal-to-noise ratio (SNR) at the destination (γ_b) exceeds the threshold SNR between source and destination (γ_{tb}). If the above consideration fails, the destination sends an acknowledgment to the source to resend the data again to the destination. In the next phase, the cooperative relay plays an important role if the communication in non-cooperative mode does not meet the threshold level. Therefore, relay operates in hybrid mode. If the instantaneous SNR between source and relay (γ_a) exceeds the threshold SNR between source and relay (γ_{ta}), then the relay decode and forwards (DF) the desired signal to the destination, otherwise the relaying protocol is getting changed to amplify-and-forward (AF) mode. Hybrid maximal ratio combining (H-MRC) technique is used to combine the signals received from the source as well as from the relay. Finally, the destination recovers original information from cooperative and non-cooperative modes.

Therefore, the signals received at the destination from the source is denoted by

$$y_b = \sqrt{P_s} h_b x + N_b \tag{3}$$

At the initial stage, the signal at the relay can be represented by

$$y_a = \sqrt{P_s} h_a x + N_a, \tag{4}$$

where h_b and h_a denotes the channel gain between SD and SR, respectively. P_s stands for power transmitted by the source. It follows Rayleigh channel distribution and x represents the transmitted signal from source. N_b and N_a are independent and identically distributed (i.i.d.) complex Gaussian random variables with mean zero and variance N_0.

Here, the system model operates in cooperative mode. Therefore, the received signal at the destination from the relay can be denoted by

$$y_c = \sqrt{P_R} h_c X + N_c, \tag{5}$$

where h_c represents the relay-destination channel coefficient. X is the transmitted signal received from the relay with respect to SNR threshold and channel quality. P_R denotes the transmission power of relay. N_c is independent and identically distributed (i.i.d.) complex Gaussian random variable. In AF mode, the signal received at the destination can be expressed as

$$y_c = \beta_r h_c y_a + N_c \tag{6}$$

In (6), the scaling factor is denoted by β_r. Therefore, the received signal would be scaled by an amplification factor β_r and it is inversely proportional to the received power, written by [4]

$$\beta_r = \frac{\sqrt{P}}{\sqrt{P|h_a|^2 + N_0}} \tag{7}$$

The above expression consists of the net amount of power and it is equivalent to the transmit power at the source. In order to obtain a better quality of service, instantaneous SNR can be utilized. Therefore, instantaneous SNR for source-destination (SD), source-relay (SR) and relay-destination (RD) terminals are represented by

$$\gamma_b = \frac{|h_b|^2 P_s}{N_0} \tag{8}$$

$$\gamma_a = \frac{|h_a|^2 P_s}{N_0} \tag{9}$$

$$\gamma_c = \frac{|h_c|^2 P}{N_0} \tag{10}$$

Equations mentioned in (8), (9) and (10) follows Rayleigh distribution. The density function of Rayleigh channel is given by

$$f_{\gamma_i}(r) = \frac{r}{\bar{\gamma}} e^{-r^2/2\sigma^2} \text{ for } r \geq 0, \tag{11}$$

where $\bar{\gamma}$ represents the average SNR of the source, relay and destination links correspondingly. Exploitation of antenna configurations in cooperative communication systems are helpful to estimate the diversity gain and coverage. Therefore, the signal has to be pre-coded prior to transmission over wireless channels. Moreover, single and multi- antennas are considered in the proposed system which uses Alamouti concept. The probability of the proposed system can be approximated as

Table 1 Simulation parameters

Parameters	Values
Block length	1024 bits
Number of relay node	1
Total power	20 dB
Channel coding	Polar coding
Modulation type	16-QAM
Antenna configuration	1×1 SISO
	2×1 MISO

$$P_e^{opt} = arg\,min_{\gamma_{ti}}\left\{f_{\gamma_i}(r)Z(W), \gamma_{ti}\right\}\sqrt{\frac{P_s}{2}}\,H_{eff}s + N, \quad i \in a, b \tag{12}$$

In (12), H_{eff} denotes the effective channel matrix and it is orthogonal. N is independent and identically distributed additive Gaussian random variable.

During the process of decoding, successive cancelation list technique is used. It consists of several survival paths to capture the erroneous information. At each decoding step, this decoder splits the decoding path into two different indices U_a and U_c. The list decoder keeps only the best paths from the above indices and rejects worst paths.

3 Simulation Results

In this section, the simulation results for the proposed polar coded incremental hybrid decode-amplify-and-forward relaying protocol are presented. The system specifications are considered over Rayleigh channel and the simulations have been carried out using MATLAB. Simulation parameters are mentioned in Table 1.

3.1 Performance Analysis

To realize the rate of reliable information transmission, bit error analysis is considered in the proposed system and was simulated using MATLAB R2017. In this section, the simulation results for polar code based IHDAF (PCIHDAF) relaying protocol are described. Communication is done through the Rayleigh channel with selective cooperation. In the simulations, the polar codes with K = 512 bits, code block length of N = 1024 bits along with Monte Carlo simulations have been used. Recent cooperative relaying schemes have been compared with the existing scheme in Table 2.

Table 2 Comparison of different relaying schemes with the proposed scheme

SNR (dB)	BER performance				
	IDF [4]	ISDF [4]	Existing IHDAF scheme [4]	Alamouti scheme [4]	Proposed IHDAF scheme
0	10^{-1}	10^{-1}	10^{-1}	10^{-1}	$\approx 10^{-2}$
10	10^{-2}	$\approx 10^{-3}$	10^{-3}	10^{-2}	10^{-4}
20	10^{-4}	$\approx 10^{-5}$	10^{-5}	10^{-4}	10^{-10}

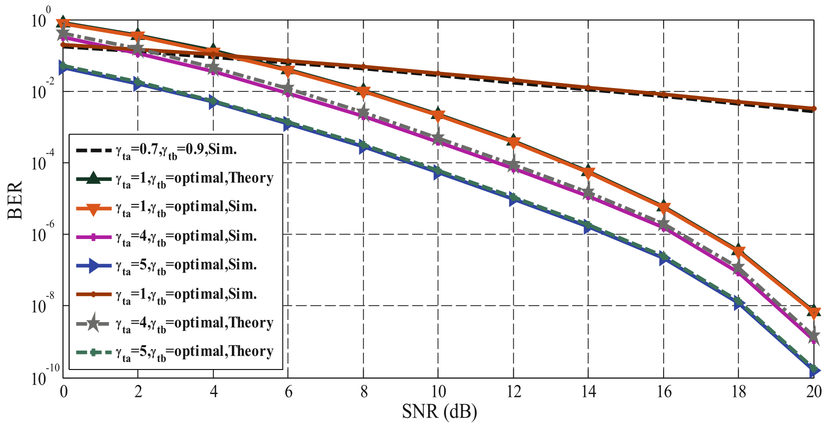


Fig. 2 Bit error rate of IHDAF relaying using polar codes for SISO system

Figure 2 shows the bit error performance for PCIHDAF relaying scheme for different values of γ_{ta} and optimal γ_{tb} with respect to total SNR of the system. From Fig. 2, it can be shown that the theoretical results are well matched with the simulation results. The simulation results show that the polar code with $\gamma_{ta} = 5$ and $\gamma_{tb} = optimal$ performs better than other values in terms of error reduction. At 18 dB SNR of the system, it provides 2 dB performance gain when compared to the low-SNR regime.

Figure 3 illustrates the error performance for incremental hybrid DAF relaying scheme through polar codes. Alamouti configuration (2×1) is considered to evaluate bit error rate. Likewise in the previous considerations, different values of threshold values of γ_{ta} and optimal threshold γ_{tb} are considered. From Fig. 3, it can be seen that the theoretical results are getting matched with the simulation results. It is inferred that the polar code with $\gamma_{ta} = 5$ and $\gamma_{tb} = optimal$ outperforms other values in terms of error minimization. It also provides 4 dB gain at 16 dB when compared to the low-SNR region, but the probability of error has been reduced as compared to SISO configurations.

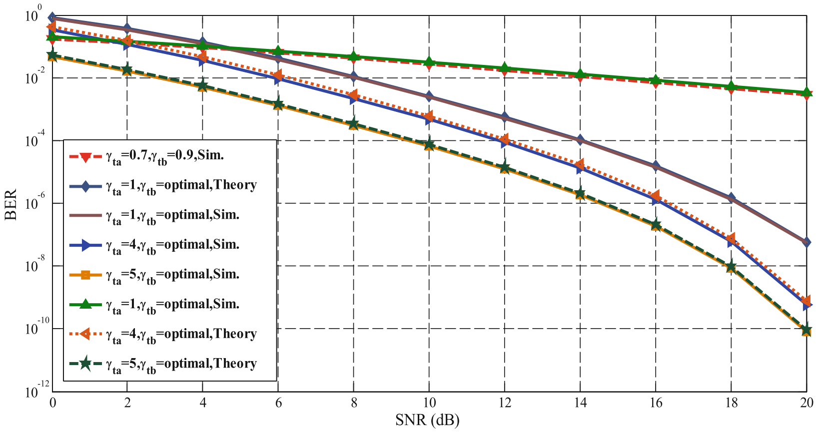


Fig. 3 Bit error rate of polar based IHDAF relaying using Alamouti scheme

4 Conclusion

In this paper, the bit error rate of the proposed polar coded incremental hybrid decode-amplify-and-forward (PCIHDAF) relaying scheme over Rayleigh channel has been investigated. The results are plotted for different threshold values of source and relay link, for SISO as well as for 2×1 antenna configurations. Simulation result shows that the proposed scheme provides the bit error rate of 10^{-10} at 20 dB signal-to-noise ratio. Monte Carlo simulations have been carried out. The future work will be the incorporation of the Genie-based relay for large MIMO systems to achieve the good quality of service in terms of capacity and coverage.

References

1. Tamer, H.M.S., Fengfan, Y., Saqib, E., Amir, A.: Decode and forward polar coding scheme for receive diversity: a relay partially perfect retransmission for half-duplex wireless relay channels. IET Commun. **11**(2), 185–191 (2017)
2. Madhusudhanan, N., Nithyanandan, L.: Compress-and-forward relaying with polar codes for LTE-A system. In: Proceedings of the IEEE International Conference on Communication and Signal Processing (IEEE ICCSP), pp. 1497–1501 (2014)
3. Venkateswari, R., Madhusudhanan, N., Ajun, A.: Bit error analysis on incremental hybrid detect-decode-and-forward cooperative relaying protocol. Natl J. Technol. **13**(4) (2017)
4. Bai, Z., Jia, J., Wang, C.X., Yuan, D.: Performance analysis of SNR-based incremental hybrid decode-amplify-forward cooperative relaying protocol. IEEE Trans. Commun. **63**(6), 2094–2106 (2015)
5. Duong, T.Q., Zepernick, H.J.: On the performance gain of hybrid decode-amplify-forward cooperative communications. EURASIP J. Wirel. Commun. Netw. (2009). Article id: 479463

6. Trung Duy, T., Kong, H.Y.: Performance analysis of hybrid decode-amplify-forward incremental relaying cooperative diversity protocol using snr-based relay selection. J. Commun. Netw. **14**(6), 703–709 (2012)
7. Ikki, S.S., Uysal, M., Ahmed, M.H: Performance analysis of incremental-relay-selection decode-and-forward technique. In: Proceedings of the IEEE International Telecommunications Conference (IEEE GLOBECOMM), pp. 1–6 (2009)
8. Liu, T., Song, L., Li, Y., Huo, Q., Jiao, B.: Performance analysis of hybrid relay selection in cooperative wireless systems. IEEE Trans. Commun. **60**(3), 779–788 (2012)
9. Niu, K., Chen, K., Lin, J., Zhang, Q.: T: polar codes: primary concepts and practical decoding algorithms. IEEE Commun. Mag. **52**(7), 192–203 (2014)
10. Zhan, Q., Du, M., Wang, Y., Zhou, F: Half-duplex relay systems based on polar code. IET Commun. **8**(4), 433–440 (2014)
11. Wang, J., Jin, A., Shi, D., Wang, L., Shen, H., Wu, D., Hu, L., Gu, L., Lu, L., Chen, Y., Wang, J., Saito, Y., Benjebbour, A., Kishiyama, Y.: Spectral efficiency improvement with 5G technologies: results from field tests. IEEE J. Sel. Areas Commun. **35**(8), 1867–1875 (2017)
12. Akuon, P., Xu, H.: Polar coded spatial modulation. IET Commun. **8**(9), 1459–1466 (2014)

A Proposed Architecture for Cold Start Recommender by Clustering Contextual Data and Social Network Data

V. R. Revathy and Anitha S. Pillai

Abstract Recommender Systems (RS) help users in selecting the apt items based on their taste from a pool of items. These systems are able to do a proper recommendation with the aid of Machine Learning algorithms. The context of a user plays an important role in recommending relevant and important product/item to a user. Social media networks are useful knowledge sources to elicit more ratings from new users than state-of-art active Learning strategies. If we are designing an RS for users whose tastes differ according to the current context (e.g., feeling), we can collect contextual data and social media information so that we will be able to recommend the right product or item. We can do this recommendation by using cross-domain RS, Selective Context Acquisition, and Implicit Feedback. This paper provides insights based on the state-of-the-art contextual data and social media environments in providing the cold-start recommendations and also propose the architecture for recommending the items to solve the cold-start issue.

Keywords Recommender systems · Cold start · Hybrid clustering · Contextual data · Latent relationship · Social network data

1 Introduction

Personalized Recommender Systems are becoming quite popular in recent days. Social Recommender Systems are developed to handle a huge amount of data in the form of number of users, and their ratings, comments, emoticon and so on in social networks [1]. On analyzing this data, it can provide valuable information. Big Data needs Machine Learning algorithms and data mining tools to tackle related challenges. Recommender systems are popular in the Big Data Analytics field with a

V. R. Revathy (✉) · A. S. Pillai
School of Computing Sciences, Hindustan Institute of Technology and Science, Chennai, India
e-mail: revathyvrajendran@gmail.com

A. S. Pillai
e-mail: anithasp@hindustanuniv.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_34

vast amount of information and also a large number of users. To produce personalized recommendations for new users, there are significant challenges. This paper attempts to explore the possibility of using social network data and contextual data to tackle new users in any given site. This information can be potentially be a challenge to fuse, but it paves way to solve the cold start problem [2, 3]. This paper, describes the design of a RS architecture incorporating Sentiment Analysis.

The personal information from the social network is useful in providing suitable recommendations to new users. This also helps in improving the prediction accuracy rate. The social media information for new users become more valuable in producing recommendations fused with contextual data [4]. This paper discusses the structure of an RS, which combines the datasets collected from the social networking sites such as *Twitter, Facebook or Pinterest* and contextual datasets of the items available from the public domains. This paper tries to study the answer for two questions:

(1) How do you recommend an item to a new user (the new user cold-start problem)?
(2) How do you recommend a new item with no ratings (the new item cold-start problem)?

The answer to the first question is the proposed design, which combines datasets from social networks and the contextual datasets from the public domains. This proposed design attempts to use the context-aware information to identify the location of the new user. Based on the current location, it clusters the items in that location and generates results. For, e.g., if it is a movie recommendation system, using the movie's contextual information, we can potentially recommend a movie either at theaters or at home. That is adjusting or filtering the recommendations using the context. Demographic social information from social networking sites like age, country, job and so on can also be used to recommend movies to a new user [3]. Extraction of meta-information about a new user can also be done from user profile information based on the user's movie ratings in the social media. The next way of eliciting user's information is asking him to sign up. Response to the second question or the new item cold-start problem is dealt with public domains like Wikipedia. The featured ratings can be collected through crowdsourcing.

The rest of this paper is structured as follows: Sect. 2 explains the recent literature on social media information and contextual data for addressing cold start problem, thus by highlighting, our findings using comparison with the proposed work. Table 1 shows how these information are useful for a personalized RS, and its remarks if we only rely on them. Section 3 describes the working of a Collaborative Filtering (CF)-based RS, on social network data and context aware data, and also presents the aspects related to the design; Sect. 4 proposes an architectural design for the cold-start recommender; Sect. 5 presents the conclusion and future enhancement.

Table 1 Comparison of similar works and proposed work

Similar works	Sources for cold-start recommendation	Techniques used	Remarks
Arain et al. [5]	Social context information and geo-location information	Probabilistic Bayesian method, Kernel density estimation, symmetric Kullback–Leibler divergence	Dependent on historical location knowledge
Sun et al. [6]	Social network information	Bi-clustering, matrix factorization	Dependent on friendship among users and ratings
Bahramian et al. [7]	User feed backs and contextual information	Artificial neural network and case-based reasoning	Not all contextual information are considered
Alahmadi et al. [8]	Social trust information	Genetic algorithm, support vector regression algorithm (SVR)	Less importance for user preferences
Jiang et al. [4]	Social context information	Probabilistic matrix factorization method	Not all contextual information are considered. Performance is based on accuracy of incorporating individual preferences and interpersonal influence
Lin et al. [9]	Social network information	Latent Dirichlet allocation	Social media site followers information is the only primary source
Reshma et al. [10]	Social network information	User-based collaborative filtering, Nearest neighbor calculation	Dependent on social media information and ratings
Levi et al. [11]	Reviews of similar contextual data	Clustering	Dependent only on similar context
Zheng et al. [2]	Contextual data, user ratings	Collaborative filtering, opinion mining	Dependent on user ratings and context data
Proposed work	User registration, social network data, contextual data, and ratings	Collaborative filtering, clustering, implicit feed back	Considers data inputs from all the additional sources and also collects additional data during user registration

2 Related Work

In the past, the cold-start problem has been studied based on additional knowledge sources or social recommender systems. This section describes related work on recommender systems which work through social network data and contextual data. Recent literature on the solutions for the cold-start problem based on social media sites is also discussed.

In [3], Revathy et al. conducted a review on the cold-start problem. This work [3] discusses different state-of-the-art approaches to solve the cold start problem through cross-domain recommender system, elicitation of user profile via applying association rules, demographic similarities, solutions via studying users' behavior on the website, user elicitation, matrix factorization and building decision trees, and applying deep learning techniques. Sassi et al. [12] conducted an extensive survey of Context-Aware Recommender Systems (CARSs) on mobile environments. They highlighted the relevance of contextual data in generating accurate mobile recommendations. This survey discusses the contextual information that represents the users' behavior. This includes personal, emotional, current activity and historical information. Authors classified the CARs into location-based, social-based, time-based, emotion-based, activity-based, and multi-dimensional-based recommender systems.

The works that are analyzed are dependent either on social networks or contextual data or social context data. In this paper, we propose a recommender system that combines both social networks and contextual data in proposing a good item to users. This is expected to minimize the cold start problem.

3 The Proposed Recommender System

The objective of the proposed recommender system incorporating Collaborative Filtering method is to produce most relevant recommendations by tackling significant challenging issues such as cold-start problem, serendipity, sparsity, lack of diversity. The proposed approach is described as follows

(1) The new user cold-start problem is handled by user ratings, context-aware data, user demographic information and sign-up process
(2) The new item cold-start problem using Wikipedia, or some other public domains.
(3) Diversity either by filtering the results or creating a cluster of items according to the rating by other users and then building the RS.

This paper proposes an architecture combining social networks and contextual data to overcome the cold start problem. This would help in designing a personalized movie recommender system by building user profiles and item profile according to the user's preferences. The Collaborative Filtering method is enhanced through clustering using contextual information. Following aspects are taken into consideration while proposing this new architecture for the movie recommender system:

(a) Requires to consider the user's interests and preferences
(b) Requires to identify the hidden relationship between user-item pairs.
(c) Requires to reveal the intention of the user's by latent feature extraction of the movies.
(d) The machine learning algorithms and metrics adopted should support system performance, for maintaining and extracting required profile information. The mining of user traits can in turn improve the performance of the system.
(e) Expect to reduce the online recommendation complexities.

4 The Proposed Architecture

This section presents a novel architecture for item recommendation by clustering the contextual information of the users, by considering the five aspects listed in the previous section. Even though the cold-start problem is highlighted throughout the architecture, there is an attempt to overcome other complexity issues. Figure 1 illustrates the architectural view of the model. As the figure shows, the proposed architecture follows the design aspects discussed above, which should be offered by an RS to generate accurate recommendations.

4.1 Components of the Proposed System

The components of the proposed system are discussed in this section. The overall process of recommending the movies using these components is shown in Fig. 2.

4.1.1 Direct/Explicit User Preference Analyzer

This component gathers the data from additional knowledge sources (social media sites and public review sites) and requests the user to register if prior user information is not available. The social networking sites are updated continuously and these updates provide demographic information of user which helps recommender system to understand user's age, hobbies and his preferences for items. Featured ratings from the review site are also taken. The signup module asks for users' preference according to his current context. Based on the data gathered, the component tries to calculate weight on each item for each user.

4.1.2 Intrinsic Preference Forecasting Through Implicit Analysis

This component essentially finds out the intention of users by mining the latent factors of the items. Unsupervised approach can be used to infer the characteristic features through the knowledge source information of a particular user and as well as specific

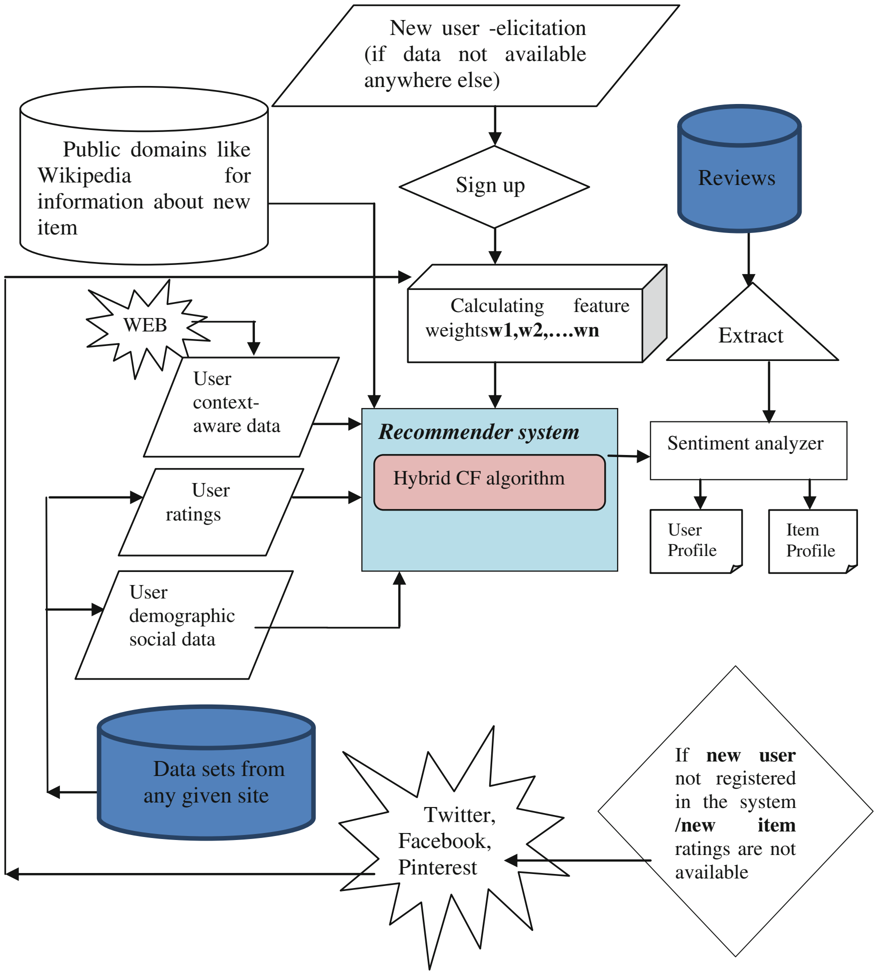


Fig. 1 Architecture of the proposed movie recommender system

information of the item. Sentiment analyzer further classifies the text reviews given by users and extracts the featured rating details. Information related to a particular item is obtained from popular knowledge sources, like Wikipedia, in which for each item, feature-value pairs exist.

This is useful in exploring an individual's feedback in the form of text reviews. The opinions or preferences about an item can be helpful in finding the contextual traits for a specific user. These traits or factors can be the cost, mood, feeling, financial status and so on, which is in turn, can find out the user's interest. This component estimates the preference score on each of the intrinsic characteristics of the item.

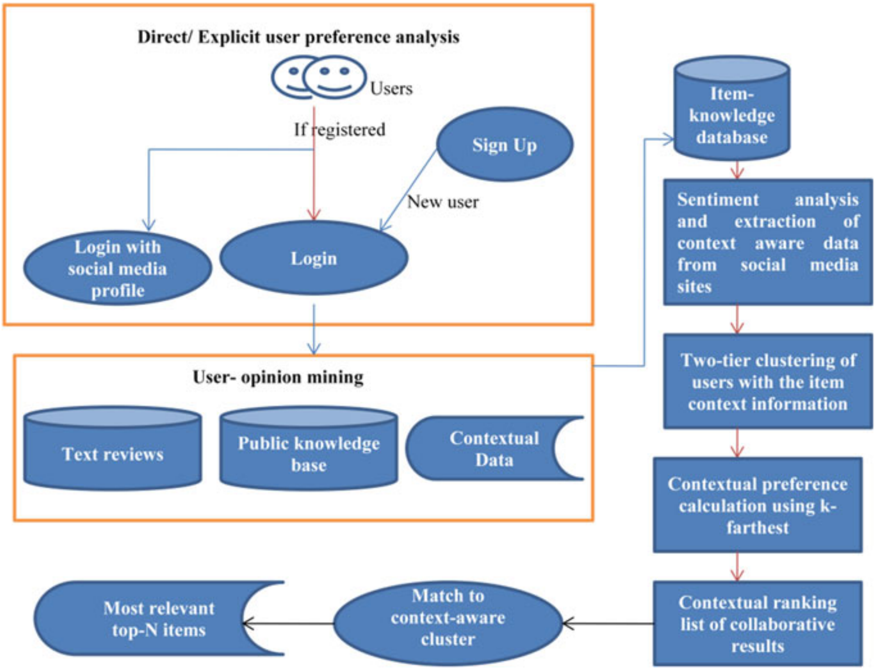


Fig. 2 The overall of process for recommending the personalized items

4.1.3 Sentiment Analyzer and Feature Extraction Through Two-Tier Clustering

This component employs a two-tier clustering approach—A. the clustering method to group the users with similar taste. The preference score of items is used to do K-means clustering for the whole users. B. The next step is to upgrade those outputs by using a sub-cluster generator and produce sub-clusters of users. This hybrid approach tries to reduce the complexities in recommending items. Along with the specific item categories and its-meta information (e.g., if it is a movie meta-information including actor, director names), those features separate clusters of users. This process helps in resolving cold start user problem by featured preferences of similar users. Integration of contextual and demographic information is the source of information enriches the new user cold-start situation.

4.1.4 Latent User-Item Relationship Recognition

The output from the two-tier clustering is used to recognize the latent relationship between a specific user and a item in the form of user-item pairs. With due importance to the positive text reviews of neighboring users in the same cluster, this component also finds out the dissimilarity among users who gave negative reviews from those existing in different clusters. This helps in enhancing the prediction accuracy rate of recommendations through essential information on personal interests. K-farthest neighbor algorithm recognizes each user according to their taste and latent feature relationship to ensure the serendipity. Moreover, the proposed approach analyzes the additional knowledge of each user regarding the current context location and his location partner. Like if L1 is location and partner is P1 "L1 P1", if the user likes an item I1 and his partner P1 also likes then "I1 P1" and if the user is alone in a location L1 then context suitable for that user only "L1 alone" to recommend the items according to the context-aware items from the list of recommendation. Both the user and item cold start problems are tackled by the two-tier clustering in terms of cluster-neighbor behavior analysis through common trait of the cluster. The extraction of meta-information for new user and new item is based on the item-related feature in user profile information and item-feature in Wikipedia information respectively.

4.1.5 Generation of Recommendation List for Each User Profile

Personalized ranking algorithm generates recommendations for the users of the proposed model. In spite of the recommended list of user–specific items, there will be a large number of overlapped similar category items. For instance, item list of a category say "C1" for the same meta-information say "I1" of that item can be suggested. This in turn, can address the inconvenience by calculating the dissimilarity between the item pairs. This approach also attempts to ensure diversity.

5 Conclusion and Future Work

Recommender systems play a crucial role in recommending either products/items to users-based on their interest/preferences. But many times it is noticed that these systems are not able to provide appropriate recommendation due to the Cold start problem. It is observed that most of the authors have attempted to solve the problem using either Social network data or Contextual data. A novel architecture combining these two methods is proposed in this paper and the various components are described. We have proposed an architecture combining the various concepts used in the papers [2, 12, 7, 11], such as clustering, contextual data, public domains, and social media to address the cold start issue.

References

1. Revathy, V.R., Pillai, A.S.: Classification and applications of social recommender systems. Advances in intelligent systems and computing. In: International Conference on Soft Computing and Pattern Recognition, pp. 719–729. Springer Cham (2016)
2. Zheng, X., Luo, Y., Xu, Z., Yu, Q., Lu, L.: Tourism destination recommender system for the cold start problem. KSII Trans. Internet Inf. Syst. **10**(7) (2016)
3. Revathy, V.R., Pillai, A.S.: Cold start problem in social recommender systems: state of art review. In: International Conference on Computer Communication and Computational Sciences (2017)
4. Jiang, M., Cui, P., Liu, R., Yang, Q., Wang, F., Zhu, W., Yang, S.: Social contextual recommendation. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 45–54. ACM (2012)
5. Arain, Q.A., Memon, H., Memon, I., Memon, M.H., Shaikh, R.A., Mangi, F.A.: Intelligent travel information platform based on location base services to predict user travel behavior from user-generated GPS traces. Int. J. Comput. Appl. **39**(3), 1–14 (2017)
6. Sun, Z., Han, L., Huang, W., Wang, X., Zeng, X., Wang, M., Yan, H.: Recommender systems based on social networks. J. Syst. Softw. **99**, 109–119 (2015)
7. Bahramian, Z., Ali Abbaspour, R., Claramunt, C.: A cold start context-aware recommender system for tour planning using artificial neural network and case based reasoning. Mob. Inf. Syst. (2017)
8. Alahmadi, D.H., Zeng, X.J.: Twitter-based recommender system to address cold-start: a genetic algorithm based trust modelling and probabilistic sentiment analysis. In: 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1045–1052. IEEE (2015)
9. Lin, J., Sugiyama, K., Kan, M.Y., Chua, T.S.: Addressing cold-start in app recommendation: latent user models constructed from twitter followers. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 283–292 (2013)
10. Reshma, R., Ambikesh, G., Thilagam, P.S.: April. Alleviating data sparsity and cold start in recommender systems using social behaviour. In: 2016 International Conference on Recent Trends in Information Technology (ICRTIT), pp. 1–8. IEEE (2016)
11. Levi, A., Mokryn, O., Diot, C., Taft, N.: Finding a needle in a haystack of reviews: cold start context-based hotel recommender system. In: Proceedings of the Sixth ACM Conference on Recommender Systems, pp. 115–122. ACM (2012)
12. Sassi, I.B., Mellouli, S., Yahia, S.B.: Context-aware recommender systems in mobile environment: On the road of future research. Inf. Syst. **72**, 27–61 (2017)

Performance Issues of Parallel, Scalable Convolutional Neural Networks in Deep Learning

Umesh Chavan and Dinesh Kulkarni

Abstract In this work, we investigate the performance issues in the parallel and scalable of Convolutional Neural Networks (CNNs). This will accelerate the training performance of CNN. In this paper we propose the parallel recognition using Compute Unified Device Architecture (CUDA) Technology and Message Passing Interface (MPI). We demonstrate scalability and performance that can be achieved on the GPU using CUDA framework where the computation-intensive tasks shifted on GPU. It compares result on GPU hardware architecture with the serial algorithm on CPU. The main novelty of our method is a new scalable CNN architecture that integrates a category hierarchy with deep CNN.

Keywords Convolutional neural network · Deep learning · MPI · CUDA

1 Introduction

Convolutional Neural Networks (CNNs) is widely applied in various research areas. For pattern recognition tasks neural networks (CNNs) work very well with large amount of training data set. The computations involve in CNN are high arithmetic intensity. With support of a new platform, the streaming model of Graphics Processing Unit (GPU), significant acceleration in training of CNN can be achieved [1]. The features of GPU make it effective to gain speedup over uniprocessor execution model (CPU), easy mapping, numerical precision as well as having strong scalability. Training phase in CNN is very compute-intensive and time-consuming task. Typically this training may take several weeks. The convolution task of CNN is well appropriate for computation on current GPUs. Paper is structured as follows: Outline of previous work is summarized in next section. CUDA model is described in Sect. 3.

U. Chavan (✉) · D. Kulkarni
Department of IT, Walchand College of Engineering, Sangli, India
e-mail: umesh.chavan@walchandsangli.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_35

Section 4 presents CNNs and explains the basic structure blocks. Implementations are explained in Sect. 5. Benchmarks and the obtained results are presented in Sect. 6. Section 7 concludes this paper.

2 Previous Work

In [2] few contributions focused their works on performance issues of CNN on the GPU. In [3] one of the first implementation of neural networks on GPU using CUDA was investigated. The goal in his work was focused on recognition part of network. Training of CNN on GPU was not considered. A fast, fully GPU implementation of CNN was presented by Ciresan et al. [4] and team. Besides this many teams are aiming on the training speed of CNN. In [5], Multi-GPU training of CNN was presented. A larger and deeper CNN with more parameters achieves better accuracy. However a larger CNN requires more computational power and training it using a typical computer is impractical [6]. The computation in a CNN can be modeled as matrix or tensor operations. Matrix operations can be easily and efficiently parallelized. Thus GPUs massively parallel processing power makes CNN to be trained efficiently and most of popular deep learning frameworks support GPU acceleration [7]. Analysis of performance of GPU-CNN framework does not exist in previous efforts. The long training time with traditional computer is not easy to meet real-time requirement in pattern recognition. Research on CNNs can be expensive and time-consuming. The process of training networks with big data sets often takes weeks on smaller computers. Using supercomputers at research institutes often require long waiting time and can be expensive. These problems led many researches to use GPUs to train CNNs, as they can be considered as high-performance parallel computers for affordable prices.

3 CUDA

NVIDIA proposed CUDA [8] (Compute Unified Device Architecture) in 2007. This created a new programming model of GPU computing and make easy and effective in using computational power of GPUs. Extending existing C program for GPU is easy in CUDA programming model. It is a C-like software environment with extensions for parallelization and special memory accesses. Other languages and application programming interfaces like FORTRAN, Direct Compute, and OpenACC are supported as well. A CUDA device is a coprocessor to the CPU and runs large number of threads in parallel. CUDA execution model involves grids, blocks, and threads [9]. Figure 1 shows the CUDA programming execution unit. A kernel in CUDA is data-parallel portion of an execution code. Kernel executes by an array of threads. All these threads run same code but different data. In the model monolithic threads are grouped into blocks. A grid is group of blocks. Thread within a block cooperates via

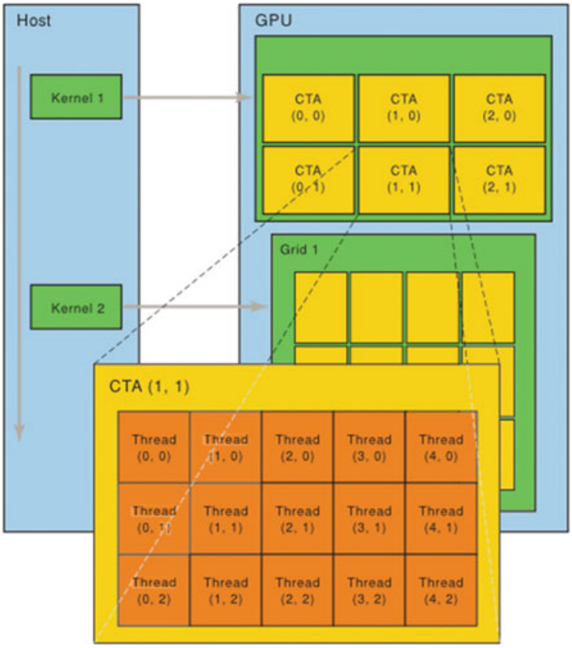


Fig. 1 CUDA architecture (adapted from [10])

shared memory. Each thread had an ID which used to calculate memory address and make control decisions. The ID is 1D, 2D, or 3D. Similarly each block has 2D or 1D ID. Group of threads physically runs on SMs (streaming processors) of GPUs. SMs share instructions. A group of threads running on SM is known as wrap. Typically wrap of 32 threads is used. The role of SMs is to schedule and allocation of threads. SM is array of streaming processors (SPs) [14].

4 Convolutional Neural Network

This structure is inspired by the receptive field structures which are found in the human primary visual cortex, first discovered by Hubel and Wiesel [6]. CNN [13] is multilayer perceptron that is design for identification of information in input data. CNN has many layers which are input layer, convolutional layer, sub-sampling layer, and output layer. The main idea of CNNs is to extract local features from input data.

It extracts relevant information (features) contents at a high resolution. The extracted features which are transition-invariant combine into more complex features. The features of CNN are expressed in terms of basic conception, which are Feature Map (FM), Weight shared, and sub-sampling [11]. Figure 2 shows, CNN arranged in a feed forward pattern. Total connection of C1 and C2 are has 1176 * (5 * 5 * 1 * 6) = 1,76,400. The architecture is represented in a structure (6,16,100,10)

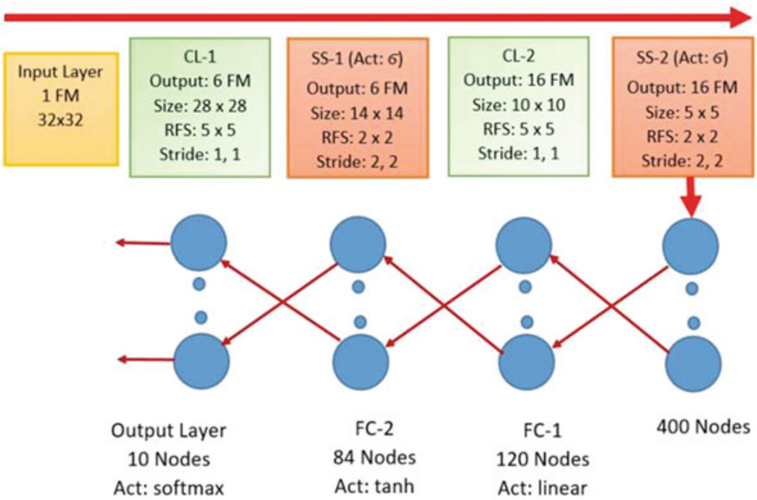


Fig. 2 CNN architecture

Table 1 Properties of LeNet CNN variations

Architecture[a]	Network properties		Trainable parameters	Speedup
	Input area	Neurons		
5-50-100-10	1,024	8,010	27,530	5
5-50-200-10	1,600	13,770	48,640	4
5-100-100-10	2,304	21,130	53,840	6
5,100,200,10	3,136	30,090	53,940	5
10-50-100-10	4,096	40,650	55020	6
10-50-200-10	5,184	52,810	195,046	5
10-100-100-10	6,400	66,570	252,646	7
10-100-200-10	7,744	81,930	319,846	7

Architecture[a] number of FMs in layer 1, 2 and number of neurons in layer 3

which represents number of FMs in CLs 1, 2 and number of neurons in FC layers [15]. The CNN represented in Fig. 2 have 8,010 neurons, 3,31,114 connections and 51,046 trainable parameters. Table 1 shows the list of CNN with different input size of data. CNN consists of major two layers—Convolutional and pooling layers.

4.1 Convolutional Layer

The Convolution is used to find the same feature in different places of an image. Convolution is conducted using learnable filters or kernels that are passed through the input data, the input image. The convolutional layer uses multiple filters where each filter moves sequentially across the input data or image to make a two-dimensional activation map based on each filter. Feature maps are made from the activation maps of the filters. And the number of learnable filters, or kernels, in the convolution process, determines how many feature maps are generated after the convolutional process.

For a two-dimensional case, the convolution can be described by the equation

$$y[m, n] = \emptyset(p) = \emptyset(b + \sum_{k=0}^{k-1} \sum_{l=0}^{k-1} v[k, l]x[l]), \tag{1}$$

where

$$\emptyset(z) = \frac{1}{(1 + \exp(-z))},$$

where k are number of kernels and l is number of layer.

4.2 Pooling Layer

Pooling layer (PL) is a sub-sampling uses a selecting operation, which is called pooling, on the feature maps. Sub-sampling is a nonlinear down-sampling process, that results in smaller feature maps. The most popular sub-sampling schemes, which there are many that exist, include median value, average value, max pooling, for each sub-region select the max value as a representative sample in Fig. 3. The output $y_n^{(l)}$ of a FM n in PL l is calculated as according to

$$y_n^{(l)} = f^{(l)}\left(w_n^{(l)} x \sum_{(i,j)\in s^{(l)}} y_n^{(l-1)}(x.s_x + i, y.s_y + j\right) + b_n^{(l)}), \tag{2}$$

where $s^{(l)}$ represents width and height of sub-sampling filter of PL l. Figure 3 illustrates this operation that is performed in pooling layer.

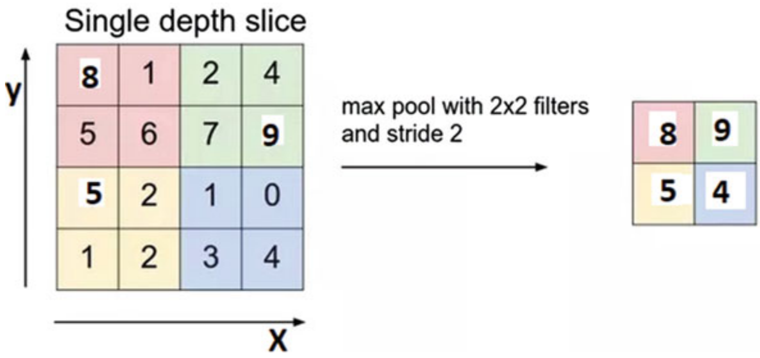


Fig. 3 Operation in pooling layer

5 Implementation

5.1 *Methodology*

We choose CNN model-LeNet and own-built model to compare the accuracy and speed of the CNN model in training and testing with GPU. We identified which part of the GPU implementation of a framework is bottleneck. The contribution of this work is as follows:

- We show the performance characteristic of the deep convolutional learning framework in GPU.
- We demonstrate the performance and scalability improvement for training on CNN to the GPU.
- We show the performance characteristics of different convolutional algorithms.

5.2 *Technical Approaches*

- GPU support

The method is explained in Sect. 3. We implemented CNN training with CUDA for GPU acceleration. The GPU implementation using CUDA exchanges the computational intensive tasks of CNN-mathematical vector and matrix operations with functions either from NVIDIA's CUBLAS library.

- Parallelizing/Distributed CNN

Table 2 GPU speedup of image of 1280×720 pixels

CPU	2.9255 s	Speedup = 105
CUDA	0.0276 s	

In the sequential implementation, calculating convolutions took more than 95% of the total training time, Parallelizing the convolution function is the best way to achieve good speedup. In convolution pixel operation is independent; it is natural to map each pixel to a thread of a block. However, in this way we will only use one block in the GPU at a time, which is a big waste of computation power. Another option is to split each image into smaller regions, and each region will be mapped to a separate block. However, this will not give us much performance gain it can even be slower than the single block solution, mainly because of false sharing. As the output image dimensions are small, without padding different multiprocessors will contend to update the output array, causing caches to be repeatedly invalidated.

6 Benchmarks and Results

The GPU we used for this experiment is GeForce GT 520 of Fermi architecture which has 48 cores. The memory bandwidth is 14.4 GB/s. GT 520 has compute capability 2.1.

6.1 GPU Acceleration for Convolution

In our first experiment we benchmarked the CPU and GPU (CUDA) convolution on scaled data size. For the experimental framework we randomly generated data sets of increasing size and observed the CPU and GPU execution time. The observation is listed in Table 2. Figure 4 shows the results which show the speedup with CUDA (GPU) version.

6.2 Scaling Feature Maps and Inner Neurons

In our second experiment in which training of CNN is compose of two phases, forward propagation, backward propagation and the weight update of CNN. We timed execution time for network for variants of CNN for different sizes of CNN architecture. The properties of CNN and speedup gain are listed in Table 1. Execution time with CPU and GPU are shown in Fig. 5. Speedup is calculated using formula.

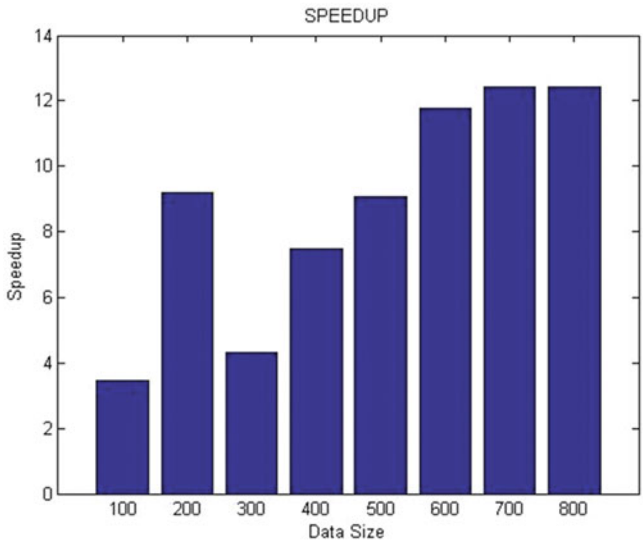


Fig. 4 Speedup between CPU and GPU for data convolution

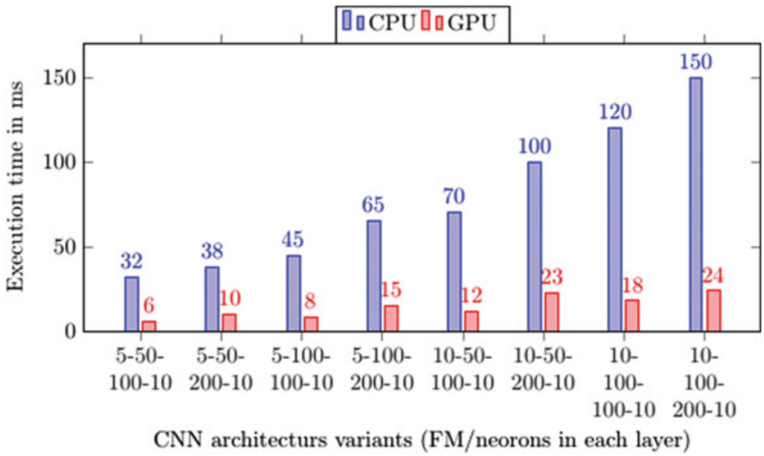


Fig. 5 Execution Time

$$\text{Speedup} = \frac{Execution\ time\ on\ CPU}{Execution\ time\ on\ GPU} \tag{3}$$

We tested the network from scaling the feature maps/neurons of convolution and FC layers. The input image size is 28×28 while the output size is fixed. Result shows that GPU variant scales better when scaling the feature maps in inner layers. The feature map is represented in format (A-B-C-D) where A indicates kernel size,

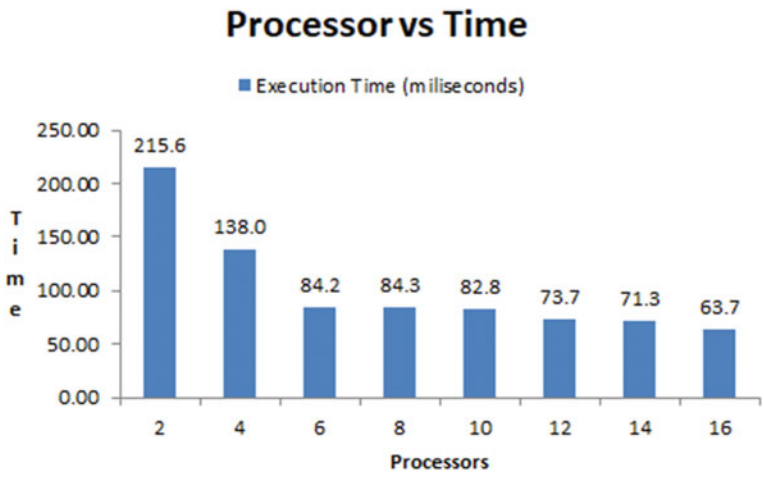


Fig. 6 Execution time with multi-process architecture

B is number of neurons in CL#1, C is number of neurons in FC layer, D is number of output.

6.3 Distributed Convolution

The program is coded using MPI [12] master-worker model. The master sends input data to a worker and worker does computations and sends back result to master. Figure 6 provides experimental results for performance of two dimensional convolutions on multi-core architecture distributed environment. It shows that convolutional algorithm is scalable.

6.4 Case Study: CUDA Image Convolution

In the last benchmarking we used CNN that is trained for speed sign detection in the image. We implemented CPU and GPU (CUDA) versions.. The convolution window has a size of 6×6. We compared the execution time of CPU and GPU version which is presented in Table 2.

7 Conclusion and Future Works

Performance improvement with GPU for CNN training is shown in this work. CUDA implementation accelerated the speedup with desired accuracy. Multi-core architecture implementation with MPI model proves that convolution is scalable operation. In the experiment which shows that high speed gain in achieved for convolution of large size image with GPU support using CUDA platform. Overall work shows that GPU/CUDA frameworks with their organization of threads are well suited for a parallel implementation of CNN. The aim of our future work is to analyze this approach to systems with multiple GPU cards.

References

1. Luebke, D.: CUDA: scalable parallel programming for high-performance scientific computing. In: Proceedings of the 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro. Paris, France [s. n.] (2008)
2. Chellapilla, K., Puri, S., Simard, P.Y.: High performance convolutional neural networks for document processing. In: Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition (2006)
3. Lahabar, S., Agrawal, P., Narayanan, P.J.: High performance pattern recognition on GPU. In: Proceedings of the National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (2008)
4. Ciresan, D.C., Meier, U., Masci, J., et al.: Flexible, high performance convolutional neural networks for image classification. In: Proceedings of the Twenty-Second In Ternational Joint Conference on Artificial Intelligence (2011)
5. Yadan, O., Adams, K., Taigman, Y., et al.: Multi-GPU Training of ConvNets. Eprint Arxiv (2013)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
7. Kim, H., Nam, H., Jung, W., Lee, J.: Performance analysis of CNN frameworks for GPUs. In: 2017 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS). IEEE (2017)
8. NVIDIA CUDA—Programming Guide, Aug 2009, pp. 55–64. http://www.nvidia.com/object/cuda_home.html
9. Kim, H., Nam, H., Jung, W., Lee, J.: Performance analysis of CNN frameworks for GPUs. In: IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Santa Rosa, CA (2017). https://doi.org/10.1109/ispass.2017.7975270
10. NVIDIA: NVIDIA CUBLAS Library. http://www.nvidia.com/object/cudahome.html. Accessed on Aug 2009
11. Lawrence, S., Giles, C.L., Tsoi, A.C.: Convolutional neural networks for face recognition. In: Proceedings of the IEEE Computer Society Conference on CVPR, San Francisco, California, USA, pp. 217–222 (1996)
12. Hedge, V., Usmani, S.: Parallel and distributed deep learning. Technical report, Stanford University, 2016. https://stanford.edu/~rezab/dao/projects_reports/hedge_usmani.pdf
13. Lam, J.C.L., Eizenman, M.: Convolutional neural networks for eye detection in remote gaze estimation systems. In: Proceedings of the International Multiconference of Engineers and Computer Scientists, vol. 1 (2008)

14. Driss, S.B., Soua, M., Kachouri, R., Akil, M.: A comparison study between MLP and convolutional neural network models for character recognition. In: SPIE Conference on Real-Time Image and Video Processing, Apr 2017
15. https://www.completegate.com/2017022864/blog/deep-machine-learning-images-lenet-alexnet-cnn/lenet-a-shallow-convolutional-neural-network

An Efficient Approach to Feature Extraction for Crowd Density Estimation

Neeta Anil Nemade and V. V. Gohokar

Abstract Crowd feature extraction is important step for crowd density estimation. This paper proposes a simple and novel approach of feature extraction applicable for crowd density estimation. A 5×5 mask is proposed for extraction of density, which finds isolated components in the image. This helps in classification of the density in five levels using SVM and ANN classifiers. The method can be used for intelligent surveillance system in public places. It can easily be used for embedded applications.

Keywords Support vector machine (SVM) · Artificial neural network (ANN) Convolution of mask · Isolated components · Statistical features · Thresholding

1 Introduction

Currently there is significant interest in visual surveillance systems for crowd density analysis. Tragedies involving large crowds occur, especially during religious, political, and musical events. To prevent the problems caused by large crowds, proper control, and management is necessary. It is convenient to know the crowd distribution from the crowd density estimation. Crowd density is one of the features of the crowd status. Crowd feature extraction is one of the key aspects of crowd analysis. Generally, there are two main targets of estimating crowd density. One is to count the number of objects approximately in the crowd and the other is to classify crowds according to the density [1, 2].

N. A. Nemade (✉)
Electronics & Telecommunication Engineering Department, SSGM College of Engineering, Shegaon, India
e-mail: neeta.nemade@gmail.com

V. V. Gohokar
Electronics & Telecommunication Engineering Department, Maharashtra Institute of Technology, Kothrud, Pune, India
e-mail: vinaya.gohokar@mitpune.edu.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_36

In recent years, many crowd density analysis methods have been proposed. The crowd density is defined for different levels as per the pedestrian per meter^2, number of pedestrian or area occupied by the pedestrians. Person counting is not always necessary for density analysis. The crowd frames are clustered according to congestion degree of the crowds. The crowd density is categorized into five levels as jammed flow, very dense, dense, restricted, and free flow [3].

First, the crowd is segmented into components using a edge detection mask. The correspondence between features extracted and the classification of the crowd density is learned. Using this relationship crowd density is categorized into five levels using the SVM Classifier.

The remainder of this paper is organized as follows; Sect. 2 presents a framework of the proposed methodology with detailed analysis of the feature extraction process. Data sets used for analysis and performance are discussed in detail in Sect. 3. Results and analysis for different datasets is done in Sect. 4.

2 Proposed Methodology

Figure 1 shows the proposed method for crowd density estimation and classification. The image acquired by the camera is first converted into gray scale for size reduction. The image is then filtered using an average filter of 5×5 mask. This pre-processing step reduces the high frequency noise components prior to the differentiation step. It helps in removal of noise as well as blurs the edges. Two types of features are extracted from the filtered image, for classification:

1. Statistical Features
2. Number of Isolated components

Various Statistical features are studied, including average, standard deviation, variance, Kurtosis. It was observed from the study on various data sets that variance is the most suitable parameter.

For detection of crowd density, a mask called K-mask is proposed which segments the crowd into number of isolated components. The mask uses principle of point detection. The purpose of mask is to detect individual people approximately. It is assumed here that, there is no occlusion. Simple edge detection or point detection mask will not be sufficient as the interest point is individual person. In order to cover larger area surrounding the pixel, the size of the mask used is 5×5. Thus, this mask is modified version of mask for point detection. The elements of the mask are both positive and negative weights. Center pixel and the nearest four neighbors are boosted while the remaining pixels of the 5×5 mask are multiplied by -2. Sum of mask weights is zero. Convolution of this mask with the background separated crowd image segments the image into individual persons. Figure 2 shows proposed K-mask and effect of convolution of this mask.

Linear special filtering of an image I (x, y) of size $M \times N$ with the K-mask is given by Eq. (1).

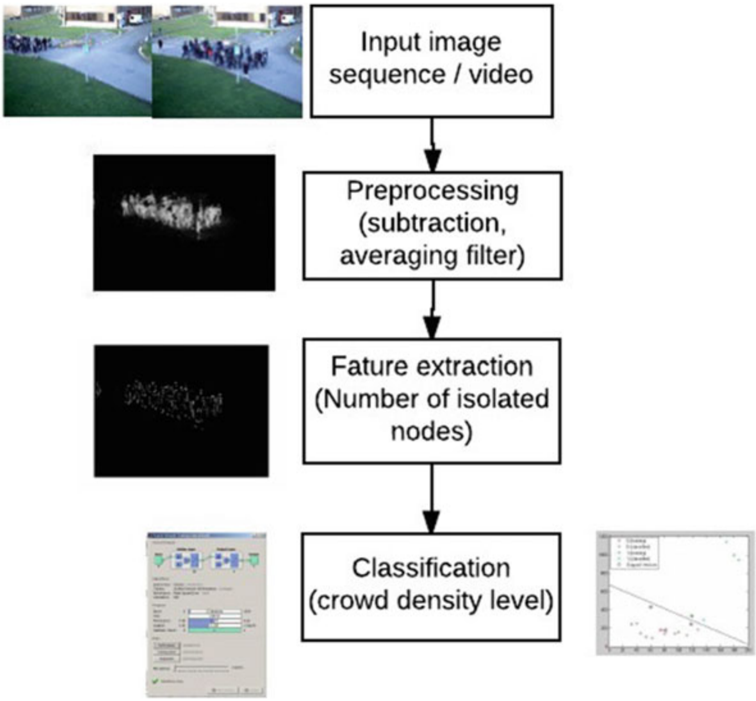


Fig. 1 Crowd density estimation flow

$$K=\begin{array}{|c|c|c|c|c|}\hline -2 & -2 & -2 & -2 & -2\\\hline -2 & -2 & 8 & -2 & -2\\\hline -2 & 8 & 8 & 8 & -2\\\hline -2 & -2 & 8 & -2 & -2\\\hline -2 & -2 & -2 & -2 & -2\\\hline\end{array}$$

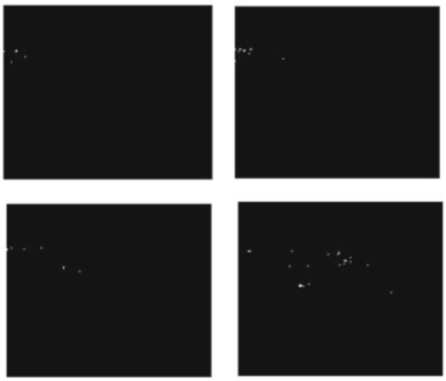


Fig. 2 k-Mask and effect of convolution of mask (segmented images) for four different crowd density levels, PETS 2009 S1, L2 walking dataset

$$g(x, y) = \sum_{s=-2}^{s=2} \sum_{t=-2}^{t=2} K(s, t) I(x + s, y + t) \tag{1}$$

The output of convolution is further applied a threshold so as to detect presence of person.

If

$$|g(x, y)| <= Threshold$$

Then

$$\mathrm{I(x,y)} = 1$$

Else

$$\mathrm{I(x,y)} = 0$$

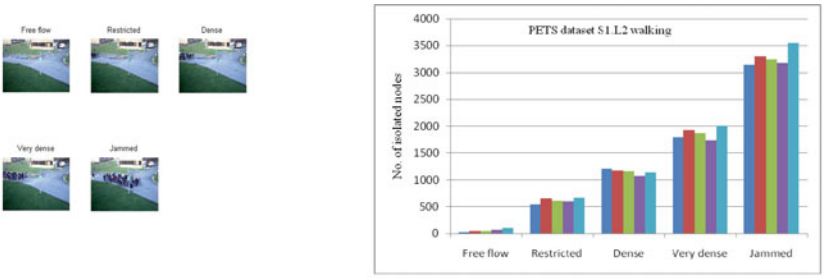
The threshold is found by averaging the minimum and maximum value of $|g(x, y)|$. The number of isolated components found from the threshold image gives an approximate value of crowd density in the image.

3 Data Sets Used for Analysis and Performance

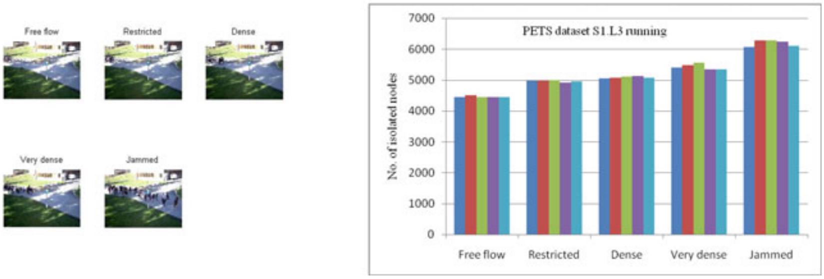
The crowd datasets used for performance are PETS 2009, UCSD, MALL and CRCV described here.

3.1 PETS Dataset (http://www.cvg.rdg.ac.uk/PETS2009/a.html)

S1 scenes are for Person Count and Density Estimation. Frames for five different crowd density levels are taken from each scenario of PETS dataset S1 for the performance. Figure 3 shows the examples for five different crowd density levels (free flow to jammed) frames for (a) L2 walking (high density crowd), and (b) L3 running (medium density crowd, bright sunshine and shadows) PETS dataset scenarios and the graph showing increasing number of isolated nodes for free flow to jammed crowd density level.



(a) Frames from PETS dataset S1.L2 walking and the number of isolated nodes for different crowd density levels



(b) Frames from PETS dataset S1.L3 running with effects of sunshine and shadow and number of isolated nodes for these crowd density levels.

Fig. 3 Frames from PETS dataset and graph between number of isolated nodes with crowd density levels

3.2 UCSD Pedestrian Dataset (http://www.svcl.ucsd.edu/projects/anomaly)

The dataset contains video of pedestrians on walkways in the campus of University of California, San Diego (UCSD), depicting crowds of varying densities taken from a stationary camera. All videos are 8-bit grayscale and the dimensions are 238 * 158 at 10 frames per second. The original video is of 740 * 480 at 30 frames per second. Figure 4 shows frames for different crowd density levels from UCSD dataset and graph between number of isolated nodes and these crowd density levels. The number of isolated nodes goes on increasing with the free flow to jammed crowd density level in this graph.

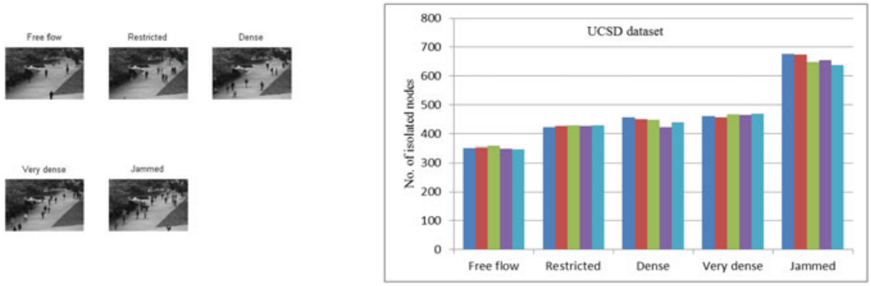


Fig. 4 Frames from UCSD Pedestrian dataset with effect of blur and the number of isolated nodes graph for 5 crowd density levels

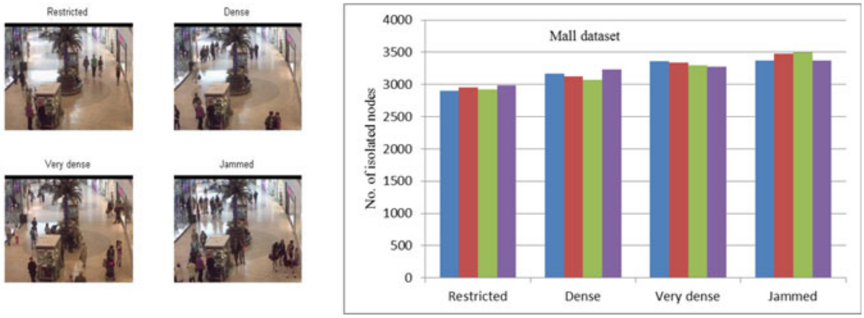


Fig. 5 Frames from mall dataset and number of isolated nodes for 5 crowd density levels

3.3 Mall Dataset (http://www.eecs.qmul.ac.uk/~ ccloy/downloads_mall_dataset.html)

The mall dataset was collected from a publicly accessible webcam for crowd counting and profiling research including video frames in jpeg format. Over 60,000 pedestrians were labeled in 2000 video frames with Video length as 2000 frames. The frame size is 640×480 with the frame rate of <2 Hz.

Diverse crowd densities from sparse to dense level and at different time of the day, different activity patterns (static and moving crowds) under larger range of illumination conditions are available in mall dataset. Figure 5 shows example frames for five different crowd density levels from Mall dataset and the increasing number of isolated nodes for free flow to jammed crowd density levels.

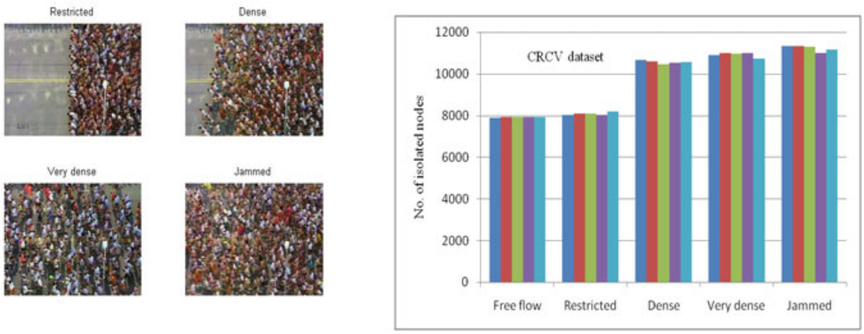


Fig. 6 Frames from CRCV (Marathon) dataset and number of isolated nodes for these crowd density levels

3.4 CRCV: Center for Research in Computer Vision (http://crcv.ucf.edu/data/crowd_counting.php)

This crowd counting dataset contains images of extremely dense crowds those are collected from the FLICKR. Example scenarios involving thousands of people are scene from New York City marathon, a large crowd participating in a political rally in Los Angeles and Pilgrims circling around Kabba in Mecca.

Marathon-1: This sequence captures participant in a marathon from an overhead camera. It is a difficult sequence due to the severe occlusion among the participants, and the similar looking outfits worn by most of the athletes. The sequence has 492 frames, but each athlete remains in the field of view, on average, for 120 frames. Figure 6 shows example frames for four different crowd density levels from CRCV (Marathon) dataset and increasing number of isolated nodes for the increasing crowd density levels.

4 Classification Results

Critical performance analysis is carried out with SVM and ANN classifiers for crowd density estimation and classification using statistical method. Good differentiation is possible between the crowd density levels with the number of isolated nodes parameter.

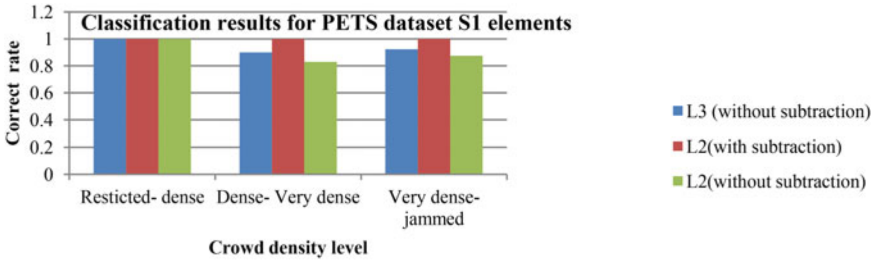


Fig. 7 SVM classification graph shows correct rate for 2 different PETS dataset

4.1 SVM Classification Results

The Support Vector Machine approach is used to classify crowd density levels.

Figure 7 shows the correct rate of estimation for two different PETS dataset with background and without background subtraction by SVM classification. Accuracy obtained from different existing algorithms is up to 80% for pixel-based (blob count), above 95% for edge-based (canny edge) and above 95% for combined foreground (background subtraction) and texture (GLCM). The margin between two crowd density levels is very important for classifying them. It is very less for restricted and dense level in blob count method, less in dense and very dense crowd level in fractal dimension but good for four levels in GLCM with subtraction method. And the margin is very good for all the five crowd density levels in proposed method [4].

4.2 ANN Classification Results

ANN classifier with back propagation network and tan sigmoid activation function shows good classification results. Fig. 8 shows results for PET dataset.

5 Conclusion

The method proposed in the paper uses a 5 $\times$ 5 mask for detection of crowd density. Convolution of this mask with the background separated image gives an appropriate measure of crowd density. Good accuracy compared to statistical parameter based methods is achieved with this approach. It is very less time-consuming. The perfor-

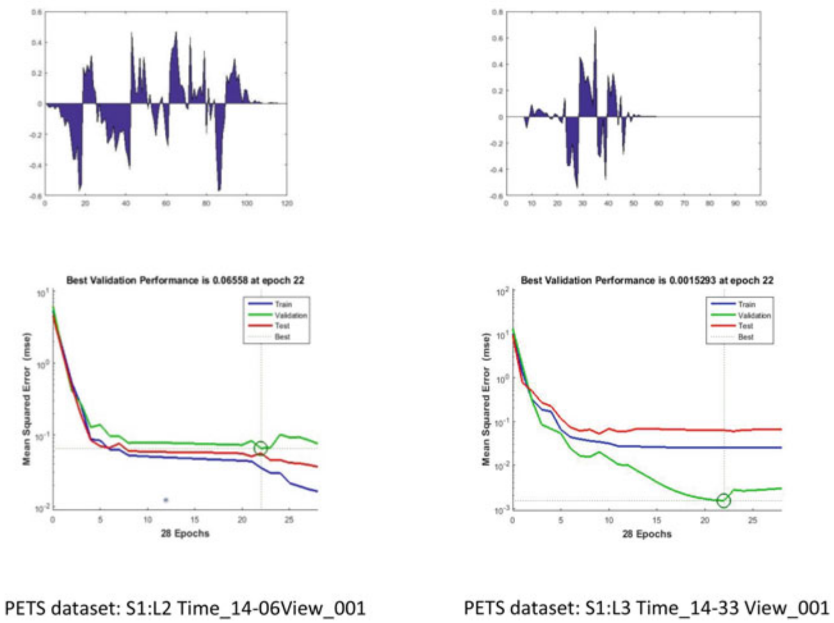


Fig. 8 ANN classification graph for PETS dataset shows network error and performance

mance parameters show effective differentiation in all five crowd density levels after testing it on state of art data sets. It is possible to implement the proposed method for real time crowd management and control.

References

1. Loy, C.C., Chen, K., Gong, S., Xiang, T.: Crowd Counting and Profiling: Methodology and Evaluation (2013)
2. Saleh, S.A.M., Suandi, S.A., Ibrahim, H.: Recent survey on crowd density estimation and counting for visual surveillance, pp. 103–114. Elsevier Ltd. (2015)
3. Polus, A., Schofer, J., Ushpiz, A.: Pedestrian flow and level of service. J. Transp. Eng. **109**(1), 46–56 (1983)
4. Nemade, N.A., Gohokar, V.V.: Feature representation for crowd counting by regression: a review. Int. J. Adv. Res. Comput. Commun. Eng. **5**(1), 402–405 (2016)
5. Nemade, N.A., Gohokar, V.V.: A survey of video datasets for crowd density estimation. In: International Conference on Global Trends in Signal Processing, Information Computing and Communication, pp. 389–395. IEEE (2016)

Book Chapter

6. Nemade, N.A., Gohokar, V.V.: Crowd Density as Dynamic Texture: Behavior Estimation and Classification, Information and Communication Technology, AISC, vol. 625, www.springer.com/us/book/9789811055072 (2016)

Unsupervised Feature Selection Using Correlation Score

Tanuja Pattanshetti and Vahida Attar

Abstract Data of huge dimensionality is generated because of wide application of technologies. Using this data for the very purpose of decision-making is greatly affected because of the curse of dimensionality as selection of all features will lead to overfitting and ignoring the relevant ones can lead to information loss. Feature selection algorithms help to overcome this problem by identifying the subset of original features by retaining relevant features and by removing the redundant ones. This paper aims to evaluate and analyze some of the most popular feature selection algorithms using different benchmarked datasets. Relief, ReliefF, and Random Forest algorithms are evaluated and analyzed in the form of combinations of different rankers and classifiers. It is observed empirically that the accuracy of the ranker and classifier varies from dataset to dataset. This paper introduces the concept of applying multivariate correlation analysis (MCA) for feature selection. From results, it can be inferred that MCA exhibits better performance over the legacy-based feature selection algorithms.

Keywords Feature selection · Supervised and unsupervised learning

1 Introduction

This sentence in nut shell implies that - Feature selection method can be used to identify subset of features which are relevant and non-redundant from a huge dataset especially which is unsupervised in nature e.g. system parameters of big data platforms (Hadoop, Spark and Storm etc.) where data cannot be labeled [1] feature selection method can be used. In case of such unsupervised and supervised data with vast attributes, it is cumbersome to analyze each and every feature. Feature selection

T. Pattanshetti (✉) · V. Attar
College of Engineering Pune, Savitribai Phule Pune University, Pune, India
e-mail: trp.comp@coep.ac.in; trpattanshetti@gmail.com

V. Attar
e-mail: vahida.comp@coep.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_37

is a technique through which the irrelevant or redundant features in a dataset can be removed making it easier to analyze the dataset with small subset of features [2]. Feature ranking is used to rank subset of features according to their significance, a similar approach wherein the attributes are ranked in accordance with the order of significance [3]. The optimal number of features, which should be selected so that the resulting subset of features is the most relevant and providing maximum accuracy remains a challenge. In order to estimate the "optimal number" of relevant features, classification algorithms are used. The subset of features, which gives maximum classification accuracy is selected. As feature-ranking algorithms are used along with classifiers for determining accuracy, the approach used here is "Wrapper" method [2, 4].

In case of labeled dataset by making use of classifier, the accuracy of feature selection algorithm can be validated but in case of unlabeled dataset, this remains a challenge [5]. To overcome this limitation, a new approach of applying multivariate correlation analysis (MCA) for feature selection is proposed. MCA can be used on both labeled and unlabeled datasets as it removes the dependency on classifying the features on labeled attributes. In this approach, correlation among features is calculated from correlation matrix. The features with higher correlation scores are chosen to be part of optimal features set. The paper is organized in following manner: In Sect. 2, discussion on existing feature selection is done. Section 3 focuses on proposed technique and its detailed working whereas in Sect. 4 the experimental evaluation is done. In Sect. 5, the work is concluded with a glimpse of future research work.

2 Literature Survey

Feature selection methods can be broadly classified into three categories namely filter, wrapper, and embedded [2, 6]. Filter approach uses the threshold measure to generate feature subset whereas wrapper approach uses prediction model by training a classifier to produce feature subsets. Wrapper method is computationally intensive because of training and testing steps involved when compared to filter. The embedded method is a collective technique wherein the feature selection is done as a part of model construction process [6].

Feature selection algorithms in filter category work on two-class problems but most real-world problems can be categorized in multiclass domain. Thus in this scenario, it is desirable that an algorithm operates in multiclass domains (as opposed to two-class domains) [7]. Hence, it becomes essential to choose algorithms that can operate on multiclass problems. While finding out relevant features it is noteworthy to consider the combined effect of several features on result. Most of the earlier algorithms do not calculate relevance as a combination of more than one feature. One of the important requirements of feature selection algorithms is that it should work on nominal as well as categorical values so as to work with real-world datasets [6]. In Relief method, relevance weight is used as a measure to distinguish the features and to classify them into two classes and features, which exceed a user-specified threshold level that is selected to form the final subset [8]. Relief [7] algorithm fails to effectively identify the relevant features in multiclass domain and to discriminate among redundant features. ReliefF is an extension of Relief and works on multiclass problem [1]. It makes use of Manhattan distance to find relevance and it gives more robust results. Methods used for calculating the distance among features can also impact feature ranking [7]. I-Relief [9] makes use of a statistical search and works on multiclass problems. RReliefF [10] is used to solve regression-based problems similar to I-Relief using statistical approach.

Fast correlation-based filter approach [11] targets the relation between feature–feature and feature–class by making use of symmetrical uncertainty coefficient measure and threshold level to restrict the number of feature and correlation measure to distinguish redundant features. Kolmogorov–Smirnov correlation-based filter test makes use of correlation measure between various features and their class labels by applying custom indexing mechanism [8, 12]. For classification- and regression-based problems Random Forest (RF) algorithm exhibits good performance [4]. However, its accuracy gets affected by high-dimensional data.

In heuristic search algorithm [13] user sets the value for identifying the number of features from the original feature set. Ensemble feature selection iteratively selects the features using data re-sampling and identify the commonly ranked features by reducing the biasing in feature ranking [14].

For preprocessing purpose, rankers are used along with classifiers to efficiently obtain the subset of features. Gain ratio and Gini do not identify the conditional dependency between the attributes optimally [15]. K-means clustering, whose space and time complexity are relatively better makes use of square error to identify the clusters. Also, it is order independent making it suitable for iterative approach [2].

In embedded unsupervised feature selection method, without making use of clustering algorithm the feature selection is embedded into the clustering algorithm through sparse learning. This has reduced the need to do transformation [16]. USFS framework is designed for feature selection of unsupervised streamed data from social media. It works effectively optimizing each step by adding new feature to the set and by removing the existing ones by making use of link information [17].

3 Proposed System

As discussed earlier most of the feature selection/ranking approaches consider each attribute as standalone feature but in real world more than one attribute contribute to achieving the final output. The relevance of features needs to be extracted by analyzing multi-feature interaction score. Based on the findings of state-of-the-art literature, the technique called multi-feature correlation score is proposed and can be used for determining the relevant features. The evaluation is carried out by comparing various feature selection algorithms with the proposed technique and the results are

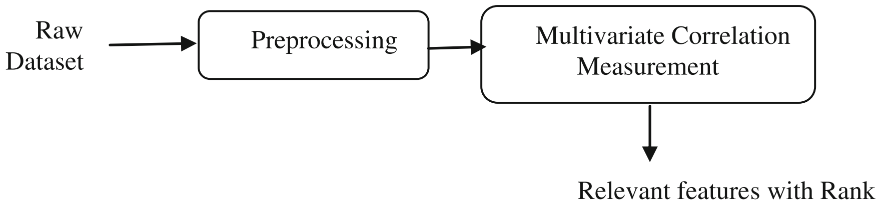


Fig. 1 System architecture

gathered over two benchmark datasets. The methodology of the proposed approach is depicted in Fig. 1.

Assume, dataset $X = \{x_1, x_2, \ldots, x_n\}$ consists of n number of samples and each sample is represented as $x_i = \{f_1, f_2, \ldots\ldots, f_m\}$. To determine the correlation of features we initially need to represent each sample as m × m matrix given in Eq. 1 in which the features are diagonal elements of matrix.

$$x_i = \begin{bmatrix} f_1 & 0 & 0 & 0 & 0 \\ 0 & f_2 & 0 & 0 & 0 \\ 0 & 0 & f_3 & 0 & 0 \\ .. & 0 & 0 & f_{m-1} & 0 \\ 0 & 0 & 0 & 0 & f_m \end{bmatrix} \tag{1}$$

$$x_i' = \begin{bmatrix} 0 & e(f_1 - f_2) & .. & e(f_1 - f_{m-1}) & e(f_1 - f_m) \\ 0 & 0 & .. & e(f_2 - f_{m-1}) & e(f_2 - f_m) \\ 0 & 0 & .. & .. & .. \\ 0 & 0 & 0 & 0 & e(fm - 1 - f_m) \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{2}$$

where e is the distance function and $e(f_i - f_j) = 0$ iff $i == j,\ also\left|e(f_i - f_j)\right| = \left|e(f_j - f_i)\right|$.

Variance between attributes is calculated using Eq. 3,

$$\sigma = \frac{1}{n-1} \sum_{i=0}^{m} \left(x'_{i,j} - \mu_{i,j}\right)\left(x'_{k,l} - \mu_{k,l}\right) \tag{3}$$

The distance between samples is calculated using Eq. 4,

$$d_{i,j} = \sqrt{\frac{\left(x'_{i,j} - \mu_{i,j}\right)'\left(x'_{i,j} - \mu_{i,j}\right)}{c}} \tag{4}$$

The covariance matrix c is computed using Eq. 5,

$$c_k = \begin{bmatrix} \sigma\left(x'_{1,1} - x'_{1,1}\right) & \left(x'_{1,1} - x'_{1,2}\right) & .. & \left(x'_{1,1} - x'_{0,m}\right) \\ \sigma\left(x'_{2,1} - x'_{1,1}\right) & \sigma\left(x'_{2,1} - x'_{1,2}\right) & .. & \sigma\left(x'_{2,1} - x'_{2,m}\right) \\ .. & .. & .. & .. \\ \sigma\left(x'_{m-1,1} - x'_{1,1}\right) & \sigma\left(x'_{m-1,0} - x'_{1,2}\right) & .. & \sigma\left(x'_{m-1,1} - x'_{m-1,m}\right) \\ \sigma\left(x'_{m,1} - x'_{1,1}\right) & \sigma\left(x'_{m,0} - x'_{1,2}\right) & .. & \sigma\left(x'_{m,m} - x'_{m,m}\right) \end{bmatrix} \tag{5}$$

Algorithm for correlation-based relevance ranking is illustrated in Algorithm 1.

Algorithm 1 Correlation Based Relevance Ranking

Input: Dataset D, number Of Samples n, number Of Features m,
Output: Correlation matrix C, Feature Importance Score R

def $\overline{\sigma}$, μ, c, sample s
S ← {Φ}
foreach $x_{k \because k \epsilon\ [1,n]} \epsilon X$ **do**
foreach i ϵ [0, m] && j ϵ [0, m] **do**
$s_{k_{i,j}} = x_{k_{i,j}}$
endFor
S ← S + $s_{k_{i,j}}$ /* add the sample into set S*/
endFor
$\mu_{i,j} = \sum_{i=0,j=0}^{i<m \&\& j<m\ (s\ \epsilon\ S)} s_{i,j} + \mu_{i,j}$ /* mean */
$\overline{\sigma}_{i,j} = \frac{\mu_{i,j}}{n}$ /* standard deviation */
foreach $s_{k \because k \epsilon\ [0,n]}$ ϵ S **do**
compute c_k using equation 5
endFor
$\overline{c} = \sum_{k=0}^{n} c_k$ /*compute average correlation score */
$R_j = \sum_{j=0,k=0}^{m} \overline{c_{j,m}}$ /*compute relevance score of each feature j */

Computed R_j is a single-dimensional matrix containing each feature relevance score. The threshold τ can be used for removing features having low-ranking score. The asymptotic time complexity of this algorithm is O(nm * (nm + 1)), where n is number of samples and m is number of features. The experimental results are presented in next section.

4 Experimental Results

The proposed technique is evaluated on publicly available datasets such as weather forecast [18] and breast cancer [19]. The dataset statistics are shown in Table 1. The results taken were compared against well-known FSA's like Relief and ReliefF. For simplicity only, a single classifier (Random Forest) was chosen for these experimentations. The relevant features extracted from the training samples are used for

Table 1 Benchmark dataset statistics

Sr. no.	Dataset	Number of samples	Number of features	Number of training samples	Number of test samples
1	Weather forecast	1648	14	1483	165
2	Breast cancer Wisconsin (Diagnostic)	569	33	480	89

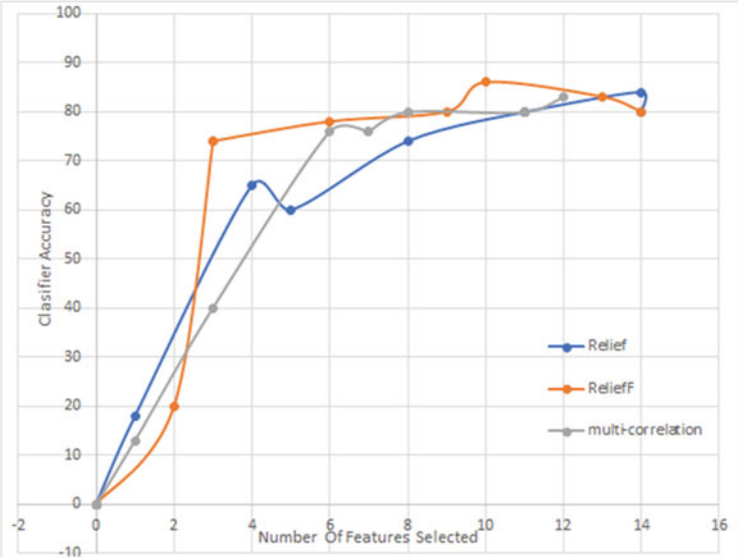


Fig. 2 Feature selection algorithm's accuracy for weather forecast dataset

training a classifier and accuracy is measured; this helps in offering a clear view of how accuracy gets impacted.

For unsupervised approach, before using the benchmark dataset for determining the accuracy of proposed feature selection algorithm, the labels were removed from the datasets and the ranked features were extracted at different threshold levels. Figures 2 and 3 show the impact of feature selection algorithm on classifier accuracy. To determine the classifier accuracy the classifier training was carried out with increasing training feature set (features were added one by one). This suggests that the ranking produced by multivariate correlation analysis holds promising prospects and can be used for further analysis and applications.

For experimentation, three feature selection approaches given as Relief, ReliefF and multivariate correlation are used. It is very obvious that different feature selection algorithms will differ in accuracy aspect for different datasets. The training samples are initially normalized to remove impurities and are then given to each of

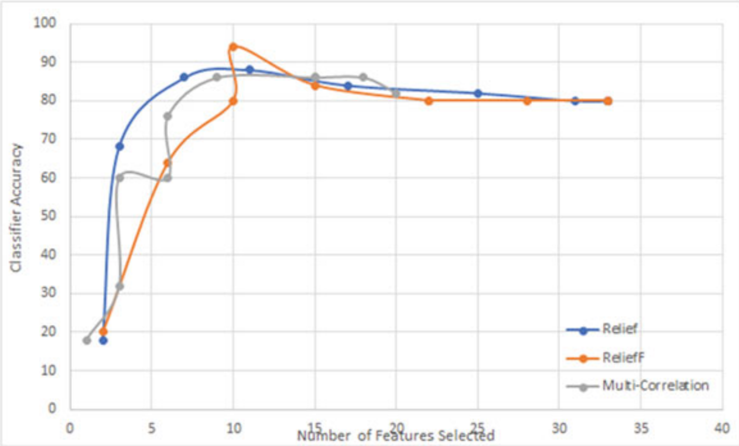


Fig. 3 Feature selection algorithm's accuracy for breast cancer dataset

the participated feature selection approaches. With every changing threshold level, the classifier (Random Forest) is trained using only selected features and accuracy is calculated. The results collected are given in Figs. 2 and 3.

The chosen breast cancer dataset is having 33 features; out of these 33 features, 10 features show high relevance. For breast cancer dataset, the maximum performance accuracy observed was 94% over ReliefF, whereas Relief and multivariate correlation approach attained nearly 86% and 90% accuracy, respectively. It can be inferred from the observation that multivariate correlation is a useful technique for feature selection.

5 Conclusion

Application of feature selection technique for unsupervised datasets with several attributes will help in identifying the irrelevant and redundant attributes. The reduced feature space can be used for real-time troubleshooting purpose. This paper proposed multi-feature correlation-based approach for feature selection. The evaluation results collected over benchmark datasets revealed a positive outcome to infer that this approach can be used for feature selection. In multivariate correlation analysis, highly correlated features are having score as +1 whereas non-correlated features are having -1 score. Features with correlation score less than 0 can be easily identified and can be removed from further inclusion. The proposed technique is useful over other techniques as it holds a promising future for unlabeled datasets. For future research work, the plan consists of applying the proposed technique to measure the accuracy over datasets consisting of hundreds and thousands of features and also to evaluate the performance of the proposed technique using evaluation metric like normalized mutual information.

References

1. Pattanshetti, T., Attar, V.: Survey of performance modeling of big data applications. In: 7th IEEE Conference on Cloud Computing, Data Science and Engineering, Confluence (2017)
2. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. 1157–82 (2003)
3. Chandrashekar, G., Sahin, F.: A Survey on Feature Selection Methods, vol. 40, pp. 16–28. Elsevier (2013)
4. Genuer, R., Poggi, J.-M., Tuleau-Malot, C.: Variable Selection using Random Forest. 31, 2225–223, (2010)
5. Mitra, P., Murthy, C., Pal, S.K.: Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Anal. Mach. Intell. **24**, 301–312 (2002)
6. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**, 273–324 (1997)
7. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: 9th International Conference on Machine Learning, pp. 249–256 (1999)
8. Gilad-Bachrach, R., Navot, A., Tishby, N.: Margin based feature selection—theory and algorithms. In: 21st International Conference on Machine Learning (2004)
9. Sun, Yijun: Iterative RELIEF for feature weighting: algorithms, theories, and applications. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 6 (2007)
10. Kononenko, I.: Estimating Attributes: Analysis and Extensions of RELIEF European Conference on Machine Learning, vol. 784, pp. 171–182(1994)
11. Yu, L., Liu, H.: Feature selection for high-dimensional data: a fast co-relation-based filter solution. In: Proceedings of the Twentieth International Conference on Machine Learning (2003)
12. Duch, W., Biesiada, J.: Feature selection for high-dimensional data: a kolmogorov-smirnov co-relation-based filter solution. Advances in Soft Computing, pp. 95–104. Springer (2005)
13. Refaeilzadeh, P., Tang, L., Liu, H.: On Comparison of Feature Selection Algorithms WS-07-05, 34-39 (2003)
14. Chi, J.: Entropy based feature evaluation and selection technique. In: Proceedings of 4th Australian Conference on Neural Networks. ACNN (1993)
15. Statnikov, A., Aliferis, C., Tsamardinos, I., Hardin, D., Levy, S.: A comprehensive evaluation of multi-category classification methods for microarray gene expression cancer diagnosis. Bioinformatics **21**, 631–643 (2005)
16. Wang, S., Tang, J., Liu, H.: Embedded Unsupervised Feature Selection, Association for the Advancement of Artificial Intelligence (2015)
17. Li, J., Hu, X., Tang, J., Liu, H.: Unsupervised Streaming Feature Selection in Social Media, CIKM'15. ACM, Melbourne, Australia (2015)
18. Weather forecast dataset link. https://nomads.ncdc.noaa.gov/data/gfsanl/
19. Breast cancer datasetlink. http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Sustainability Assessment by Use of Fuzzy Logic—A Review

Pratibha R. Dumane, Anuja D. Sarate and Satishkumar S. Chavan

Abstract This paper discusses the use of fuzzy logic for assessment of sustainability. The paper also reviews the social, economic, and environmental factors on which sustainability of the current and future generations is based. Barriers towards achieving sustainability are enlisted. The major objective of this paper is to review Sustainability Assessment by Fuzzy Evaluation (SAFE) model for assessment of sustainability. As fuzzy logic is able to deal with data that is not well defined, ambiguous and also there is a complex relationship between the parameters involved, it is well suited to the assessment of sustainability. The SAFE model is useful for policy makers to decide on the measures to be taken for sustainable development in the years to come for a sustainable future for life on planet earth.

Keywords Sustainability · Fuzzy logic · Sustainability assessment
Sustainability indicators

1 Introduction to Sustainability

For life to be sustainable on earth, it is important that the current generation uses resources wisely so that their availability is also ensured for well-being of the future generations. Human well-being depends on nature, where there is harmony and only when humans follow the laws of nature. For example, energy is required for the basic needs and services and therefore it is a very important parameter for economic growth [1]. The rate of energy consumption has increased day by day which resulted

P. R. Dumane · A. D. Sarate · S. S. Chavan (✉)
Department of Electronics and Telecommunication Engineering, Don Bosco Institute of Technology, Kurla (W), Mumbai 400070, Maharashtra, India
e-mail: satyachavan@yahoo.co.in

P. R. Dumane
e-mail: pratibhadumane@gmail.com

A. D. Sarate
e-mail: anuja.sarate13@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_38

into greenhouse gas emissions. This has adverse effect on sustainability [2]. If policy makers needed to identify potential industrial areas, there is a need to identify potential zones [3] by studying the sustainability not only from a current perspective but for the future too.

Sustainability can be broadly considered in terms of social, economic, and environmental factors. A proper balance of all these three factors is a must to ensure sustainability. With issues like global warming, ozone depletion, health, education, etc., now the policy makers need to put in efforts towards sustainable development.

In more general terms, sustainable development is defined as the growth which fulfills the needs of people without compromising over next generation ability to meet their own needs as per the Brundtland Report [4]. It is also described as the growth that provides the comfort to human life with existing abilities of ecosystems according to International Union for Conservation of Nature (IUCN).

Sustainability needs to take into consideration the time scale along with the other factors mentioned. As policies and goals are defined for the current generation, a vision for the future cannot be ignored. Sustainability needs to consider the integration of social, economic and environmental factors; none of the factors can be given importance at the cost of the other two. While policy makers can debate over the factors to be considered, most of the factors can be categorized under the social, economic and environmental parameters.

This paper discusses about the barriers in achieving sustainability (Sect. 2), scientific approaches to measure sustainability (Sect. 3), and SAFE model (Sect. 4). The concluding remarks are made in Sect. 5.

2 Barriers Towards Achieving Sustainability

Economists may give more importance to economic growth as compared to other factors for sustainable development. They may consider environment as a part of the economy. The natural resources bestowed upon all the countries are different across the globe. A proper utilization of these resources at present and ensuring their availability for the future needs to be given notable importance [5]. And this will happen if these resources are scientifically studied and looked into using appropriate mathematical models. A transfer of knowledge between researchers, scientists, and policy makers will help in assessing models in sustainability. Social barriers e.g. population growth, lack of awareness can play an important role in maintaining sustainability [6]. Implementation of various policies and projects at all the governmental levels also can be a major barrier to sustainable development [7].

3 Measuring Sustainability

Sustainability assessment is a big challenge to researchers owing to the number of parameters involved along with their inter dependence. There have been many approaches till date [4]. Some of the scientific approaches are outlined as follows:

(a) Pressure-State-Response (PSR): PSR model focuses primarily on the environmental aspects. The Organization for Economic Co-operation and Development (OECD) has developed this model of sustainability [8]. It also looks into social and economic indicators.
(b) Ecological Footprint: Rees [9] has developed this model. It mainly focuses on ecological side.
(c) Barometer of Sustainability: International Union for the Conservation of Nature (IUCN) has introduced a visual tool of sustainability assessment [10] which is known as Barometer of Sustainability. Ecosystem well-being and Human well-being are two fundamental components looked after by this model.
(d) Environmental Sustainability Index (ESI): It is an environmental sustainability index which is based on 21 indicators with 76 data sets [11]. It is computed as an index for a country.
(e) Sustainability Assessment by Fuzzy Evaluation (SAFE): Objective assessment of sustainability is modeled using Sustainability assessment by fuzzy evaluation (SAFE). It uses fuzzy logic based analysis with the help of basic indicators namely economic efficiency and social welfare with environmental integrity. It also provides estimate of human, ecological, and overall sustainability. It was presented by Phillis and Andriantiatsaholiniaina [12] and modified by Andriantiatsaholiniaina et al. [13], Kouloumpis et al. [14], and Phillis et al. [15]. SAFE structure has hierarchical inference. It utilizes 75 basic indicators which describes various environmental and societal aspects. The overall sustainability index in estimated in the range of {0–1}.
(f) Multiple-Criteria and Fuzzy Logic: Liu has introduced a model based on 74 basic indicators and multiple-criteria decision-making (MCDM). It is associated with fuzzy based inference scheme [16]. The average sustainability index is estimated using sequential fuzzy logic scheme. The MCDM utilizes analysis, biasing, and synthesis.
(g) Sustainable Society Index (SSI): This model used classification 22 indicators into 5 categories with equal weight. These indicators are based on environmental and societal parameters [17]. These five classes are aggregated into numerical measure of sustainability as SSI. It has been accepted by 150 countries for ranking.

4 An Overview of SAFE Model

Fuzzy logic discriminates the information such as vague, ambiguous, imprecise, noisy, or lost data, emulating approaches of skilled people, and supporting the human in complex mathematical analysis. It becomes a very useful tool while analyzing sustainability as it allows the experts to handle data which is uncertain and sometimes may be ambiguous too [2].

Sustainability indicators help in resolving the information for better analysis supporting in achieving objectives like preserving biodiversity, social welfare, economic performance, and balanced environmental. The SAFE model is used for assessment of urban sustainability [18].

Ecological sustainability (ECOS) and Societal or human sustainability (HUMS) are two major components of the overall sustainability (OSUS) of a country. These parameters are given in Fig. 1.

Land integrity (LAND), Air Quality (AIR), Water Quality (WATER), Biodiversity (BIOD) are the subsections of ECOS [19], whereas HUMS provides four secondary sections as Political aspects (POLICY), Health (HEALTH), Economic wel-

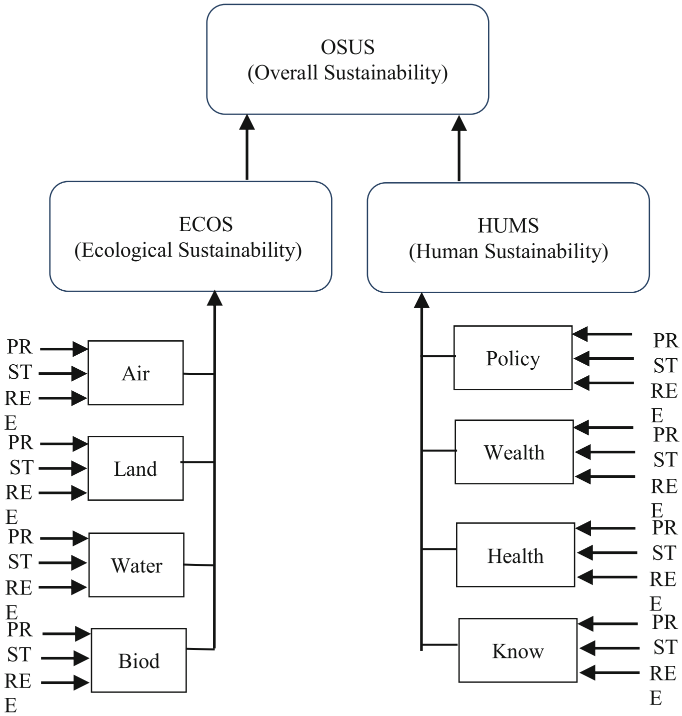


Fig. 1 Overall Sustainability (OSUS) Structure

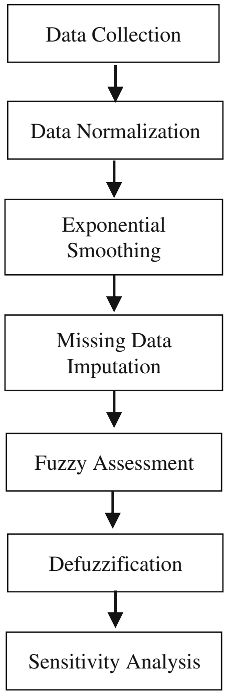


Fig. 2 Sensitivity analysis using fuzzy logic

fare (WEALTH), Education (KNOW) as presented in Fig. 1 for overall sustainability (OSUS) model.

To assess the secondary sections of sustainability model, Pressure-State-Response (PSR) approach of the Organization is used [18]. It also considers that the humans are responsible for the pressure on the environment which modify the conditions (states) of environment. It results into peculiar responses by the societal interconnections.

Pressure (PR), State (ST) and Response (RE) are tertiary indicators in the SAFE model. These indicators are used to compute the effects on biodiversity. The percentage effect on various threatened species is indicated with the help of six sustainability indicators.

In SAFE model, data processing is done mainly in sequential steps as shown in Fig. 2. The steps are explained below as

1. Data Collection and Normalization: Normalization is done by representing least desirable indicator with value 0 and value 1 is used to indicate the most desirable indicator. This is necessary to make the indicators comparable. Experts, laws and standards determine the desired and non-desired indicators that will affect sustainability.
2. Exponential Smoothing: To deal with issues where the data is imprecise or unavailable or where there is significant cumulative effect of previous data, the present and past indicator data is combined into a single value using exponen-

tially weighted sums. ECOS and HUMS both depend on the current data as well as the past data and as a result need memory. There is a need to consider past data too while evaluating the sustainability assessment. A method of exponential smoothing is used to take account of the past and via exponential weighting these past values are factored [14, 20].

3. Missing Data Imputation: Fuzzy logic helps in imputation of lost data which may not be available for few countries. These values are attributed using fuzzy logic system.
4. Fuzzy Assessment: Using fuzzy logic tool, the system can be modeled without detailed mathematical descriptions, using qualitative as well as quantitative data. This is especially useful when the data is often ill - defined because of the underlying issues. Composite indicators are found out from the basic ones. Fuzzification is used to find out the degree to which a basic indicator belongs to a specific fuzzy set. The fuzzy rules express the interdependence of the different parameters.
5. Defuzzification: Fuzzy statements can be converted into a single crisp value of overall sustainability. Defuzzification is carried out using the center-of-gravity formula.
6. Sensitivity Analysis: Decision makers can then reflect upon the most important indicators. Different scenarios for development after assessing the sustainability can be worked out. This comprises of the computation of the OSUS, HUMS, and ECOS gradients as per each and every basic indicator. A gradient will give the sustainability per unit rise of some basic indicator.

Total 75 sustainability indicators are used to get a clear picture of inter dependence and long term implications of the decisions taken for 128 countries based on these indicators [14].

For sustainability, human development as well as protection of environment is important. Green Climate Fund (GCF) is a financial instrument to assist developing countries in moderating emissions and adjusting to impact of climate change. The major developing countries including India are demanding early capitalization of GCF. It will fetch intellectual property rights (IPR) for such expensive green methodologies [21]. Table 1 gives summary of focus of the various models that are used by researchers.

Table 1 Various models of sustainability assessment and their focus

Model	Focus
Pressure-state-response	Environmental aspects
Ecological footprint	Ecology
Barometer of sustainability	Ecosystem and human well-being
Environmental sustainability Index	Environment
Sustainability assessment by fuzzy evaluation	Ecology, human and overall sustainability
Sustainable society index	Environmental and societal parameters

It is important that all the developed countries come forward to support the sustainability analysis so that appropriate measures can be taken. We cannot have human development at the cost of environment. The models will help in studying the effects of current and future actions so that policies can be framed once the outcome of the actions can be predicted.

Sensitivity analysis is one good method for managing sustainability. The results of sensitivity analysis help in taking the right precautionary measures at the right time.

5 Conclusion

The paper reviewed the use of SAFE model for assessment of sustainability. The use of fuzzy logic in SAFE model makes it simple and is the only approach that considers the time dimension. This will help in analyzing the effects of the highly inter linked parameters on the social, economic and environmental front and will be a good tool to forecast the effects of the various parameters for a sustainable development to find the problem areas and fix them before the situation worsens. Growth in any field in any part of the world cannot be at the cost of sustainability.

Assessment of sustainability involves social acceptance which further involves subjectivity and uncertainty. Fuzzy logic and fuzzy theory deal with such subjective and uncertain information which may sometimes be ambiguous. Hence, fuzzy logic becomes most helpful tool for selecting indicators for assessing sustainability. Assessment is bound to vary based on sustainability indicators chosen. More the number of indicators better will be the model. The type of fuzzy model, selection of membership function, fuzzy decision-making, defuzzification technique used will have an impact on the assessment of sustainability.

More focus needs to be given to also include emerging parameters while doing the multi-criteria evaluation for sustainability.

References

1. Brenna, M., Foiadelli, F., Roscia, M.: Sustainability energy indicators by means of fuzzy logic. Power Engineering Society General Meeting, pp. 1–7. IEEE (2007)
2. St Flour, P.O., Makoondlall-Chadee, T., Bokhoree, C., Mohee, R.: Structured fuzzy based methodological approach towards sustainability performance assessment. Int. J. Environ. Sci. Dev. **5**(2), 223–227 (2014)
3. Puente, M.C.R., Fernandez Diego, I., Maria, J.J.O.S., Hernando, M.A.P., de Arroyabe Hernaez, P.F.: The development of a new methodology based on GIS and fuzzy logic to locate sustainable industrial areas. In: 10th AGILE International Conference on Geographic Information Science, Aalborg University, Denmark, pp. 1–8 (2007)
4. Sotoudeh, M.: Links between sustainability and technology development. IEEE Technol. Soc. Mag. 9–14 (2005)

5. Bond, S., Perrett, G.: The key drivers and barriers to the sustainable development of commercial property in New Zealand. J. Sustain. Real Estate **4**(1), 44–77 (2012)
6. Jespersen, J.: The Barriers to Sustainable Development. Insight (2012)
7. Reid, M., Schwab, W.: Barriers to sustainable Development-Jordan's sustainable tourism strategy. J. Asian African Stud. **41**(5/6), 439–457 (2006)
8. Linster, M.: OECD Environmental Indicatiors—Development, Measurement and Use, Reference Paper by Organization for Economic Co-operation and Development, pp. 1–37 (2003)
9. William, E.: Rees: ecological footprints and appropriated carrying capacity: what urban economics leaves out. Environ. Urban. **4**(2), 121–130 (1992)
10. Robert Prescott-Allen: Barometer of Sustainability—What it's for and how to use it, A booklet, International Union for the Conservation of Nature (IUCN), pp. 1–25 (1996)
11. Esty, D.C., Levy, M., Srebotnjak, T., de Sherbinin, A.: In: 2005 Environmental Sustainability Index: Benchmarking National Environmental Stewardship, Chapter 2, pp. 11–19. Yale Center for Environmental Law & Policy, New Haven (2005)
12. Phillis, Y.A., Andriantiatsaholiniaina, L.: Sustainability: an ill-defined concept and its assessment using fuzzy logic. Ecol. Econ. **37**(3), 435–456 (2001)
13. Andriantiatsaholiniaina, L., Kouikoglou, V.S., Phillis, Y.A.: Evaluating strategies for sustainable development: fuzzy logic reasoning and sensitivity analysis. Ecol. Econ. **48**(2), 149–172 (2004)
14. Kouloumpis, V.D., Kouikoglou, V.S., Phillis, Y.A.: Sustainability assessment of nations and related decision making using fuzzy logic. IEEE Syst. J. **2**(2), 224–236 (2008)
15. Phillis, Y.A., Kouikoglou, V.S.: Sustainability Indicators, Fuzzy Measurement of Sustainability, Chapter 3, pp. 65–72. Nova Science Publishers, New York (2009)
16. Liu, K.F.R.: Evaluating environmental sustainability: an integration of multiple-criteria decision-making and fuzzy logic. Environ. Manag. **39**(5), 721–736 (2007)
17. van de Kerk, G., Manuel, A.R.: A comprehensive index for a sustainable society: the SSI—the sustainable society index. Ecol. Econ. **66**(2–3), 228–242 (2008)
18. Hincu, D.: Modelling the Urban sustainable development by using fuzzy sets. Theor. Empir. Res. Urban Manag. **6**(2), 88–103 (2011)
19. Pislaru, M., Trandabat, A., Schreiner, C., Spiridonica, A.: Fuzzy model for sustainability assurance related to environmental protection. In: 10th International Conference on Development and Application Systems, Romania, pp. 116–118 (2010)
20. Hsueh, S.-L.: A fuzzy utility-based multi-criteria model for evaluating households' energy conservation performance: a Taiwanese case study. Energies **5**, 2819–2834 (2012)
21. Times of India, Nov 16, p. 17 (2014)

Sentence Level Sentiment Identification and Calculation from News Articles Using Machine Learning Techniques

Vishal S. Shirsat, Rajkumar S. Jagdale and Sachin N. Deshmukh

Abstract Sentiment analysis is a widely used phenomenon for analyzing online user responses to infer collective response and it is used in various applications. Negation is a very common morphological creation that affects polarity. This research paper focuses on sentence level negation identification from news articles this work uses online news articles Data from BBC news. Results are analyzed using Machine Learning Algorithms like Support vector Machine and Naïve Bayes. Support Vector Machine achieves 96.46% accuracy and Naive Bayes achieves 94.16%.

Keywords Sentiment analysis · Support vector machine · Naïve Bayes · Machine learning algorithm · Negation identification

1 Introduction

Sentiment Analysis is an application of Natural Language Processing, computational linguistics and text analytics which identifies and extract subjective information in source materials such as product reviews, chats and discussions [1]. Sentiment analysis determines the inclination of a correspondent through the contextual polarity of their language or writing, their attitude which may be pretended in their own judgment, emotional state of the substance, and otherwise the state of any emotional communication they are using to affect a reader. It is demanding to define a person's state of mind on the topic they are collaborating about. This information can be mined from several data sources from texts, tweets, blogs, social media, and news

V. S. Shirsat (✉) · R. S. Jagdale · S. N. Deshmukh
Department of Computer Science and IT, B. A. Marathwada University, Aurangabad, India
e-mail: vss.csit@gmail.com

R. S. Jagdale
e-mail: rajkumarjagdale@gmail.com

S. N. Deshmukh
e-mail: sndeshmukh@hotmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_39

articles [2]. News articles and web blogs are one of the most essential platforms that permit users to express their personal opinion about the several topics. Basically Sentiment analysis covers a big part of, computational linguistics, natural language processing, and text mining. Generally, the aim of sentiment analysis is to finding the polarity of opinion. In statistical way, sentiment analysis methods are based on frequency of positive and negative words. Many researchers have identified the ways of accounting for several other features of content, for example structural aspects.

2 Related Work

Nowadays online news and web blogs have become an important source of information from news websites. People share their thoughts, feelings in the form of news articles and web blogs. This increase in the amount of online opinion-related textual information has led to the rapid development of the field of sentiment analysis. Pang et al. [3] have used machine learning approaches to determine the accuracy of classification from documents. Experiments were performed on movie data and it was concluded that the machine learning techniques are always better than human made baseline for sentiment analysis. The work involves intensifiers analysis to extract exact sentiments. Machine Learning approaches like Naïve Bayes, Maximum entropy and support vector machines classification techniques are used. As an inference, it is concluded that machine learning techniques are better than human baselines for sentiment classification. Mohammad et al. [4] defined a method to increase the scope of sentiment lexicon and includes the Identification of separate words and multi-word expressions. Lexicon and a list of affixes are used here. This method can be implemented using antonym generation or lexicon based. For antonym generation Hand-crafted rules were used. Lexicon approach is based on the word list which defines if a paragraph has more negative words than the positive ones and accordingly polarity of the paragraph is decided. Turney [5] has used semantic mining for binary classification and part of speech (POS) tagging. He worked on document level and review level sentiment analysis. Shoukry [6] shows an application for Arabic language sentiment analysis and performed a sentiment classification for Arabic tweets. The collected tweets are examined to provide their polarity. Their study proposed hybrid system that used all the identified features from the ML approach, and the sentiment lexicon from the SO approach, resulting in an accuracy and recall of 80.9%, while its precision and F-measure is 80.6%. Alexandra Balahur [7] stated the importance of the tasks in three levels. The work separated good and bad news content. Ding and Melville [8, 9] focused on machine learning approaches to train classifier. Lexicon dictionary based method depends on corpus or list of words having certain polarity. An algorithm pointed out the dictionary words and calculated the weight accordingly.

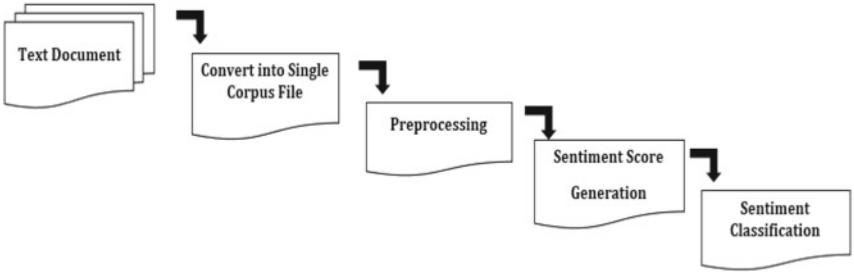


Fig. 1 Proposed methodology

3 Preprocessing

Sentiment analysis is the process which identifies or expresses the polarity of the text data. Basically, Sentiment analysis has been categorized in three levels: first Document level sentiment analysis, here whole document is to be to find the polarity of that document. For example, if one text file contains reviews of only one product, then the system calculates polarity of complete text in the document. Thus the document expresses opinion on a single entity and is not applicable on multiple product reviews [10]. In sentence level, every sentence is processed and analyzed to determine the polarity. Aspect level sentiment analysis helps to realize find out sentiment on objects and their features [11].

The preprocessing of the dataset and preparing the text for classification is an important task. The work presented in the paper uses BBC News Article Dataset. Online text contains irrelevant text such as HTML tags, scripts, and advertisements. Preprocessing plays a very vital role in text mining methods and applications. It is a first and foremost steps which will help to cleaning a data and increasing the data sparsity and substantially shrinking the feature space [12]. There is no impact on the general orientation in word level sentiment analysis [13]. It also helps for to cleaning and preparing the text for classification. Basically online and offline data is having huge amount unwanted information which does not contain any wanted information from the data. Removing the unwanted information and meaningless information, from the text data such type of things comes under in preprocessing stages (Fig. 1).

4 Machine Learning Algorithms

Naïve Bayes classifier uses Bayes Theorem, which finds the probability of an occurrence given the probability of another occurrence that has already occurred. NB classifier does particularly well for problems which are linearly separable and even for problems which are nonlinearly separable it perform reasonably well [14].

$$P(A|B)\frac{P(B|A)P(A)}{P(B)} \tag{1}$$

Support vector machine is non probabilistic algorithm which is used to separate data sequentially and Non-sequentially [15]. It is basically used for text classification and get a good performance in high-dimensional feature space. Support Vector Machine algorithm denotes of the instances points in space, mapped so that the instances of the different classes are separated by a clear margin as extensive as possible [16].

5 Experimental Results

Our proposed methodology uses mainly five steps. In first steps performs data cleaning and removes URL, Stop words, Punctuation, Strip white space and Numbers from the data. The step of number removal is important a number hardly represents the sentiments and hence not useful. Then the next step performed is to convert the whole document into lower case to have uniformity and then the stemming. Stemming is used for to change word a root form of the word for example. "Education", "Educated", "Educating" will be converted into single word, i.e., educate. After preprocessing the dataset, we need to determine the Term Document Matrix which describes the frequency of terms that occur in the processed dataset. Rows in the dataset are considered as a collection and column considers as a related terms. This is achieved by using "dtm" function in TM package of R. After preprocessing, there is a need of sentiment score generation with the help of positive and negative dictionary. Each word in the dataset will be compared with the dictionary word to determine whether it is positive or negative. Further, Naïve Bayes and Support Vector Machine algorithms are used for the classification purpose and accuracy is estimated.

Table 1 below shows the category of the article and the count of neutral, positive and negative word in it. This work uses Bing Liu dictionary which contains 2006 positive word and 4783 negative word. The results are as following (Fig. 2).

For the classification two machine learning algorithms Naïve Bayes and Support Vector Machine are used, and Accuracy, Precision and F-Score of the Data is

Table 1 Category wise document polarity

Sr No	Name of category	Total	Positive	Negative	Neutral
1	Business	510	262	214	34
2	Entertainment	401	136	244	21
3	Politics	417	210	190	17
4	Sport	511	151	327	33
5	Tech	401	136	244	21

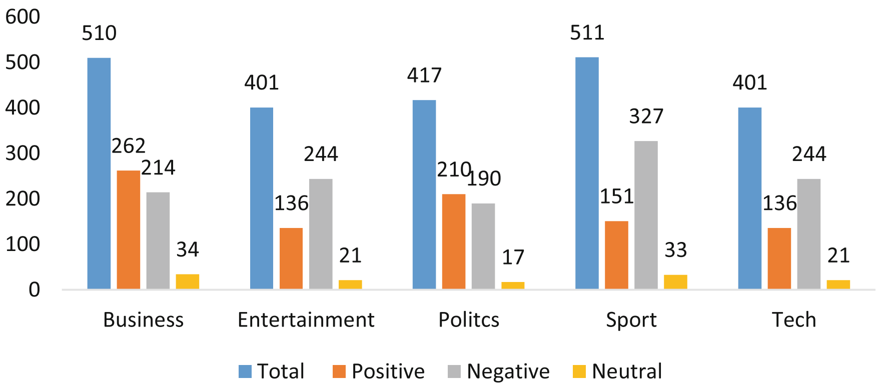


Fig. 2 Graphical representation of document polarity

Table 2 Comparative analysis of Naïve Bayes and SVM

Dataset	Naïve Bayes			SVM		
	Accuracy	Precision	F-score	Accuracy	Precision	F-score
Business	92.63	89.76	91.32	82.60	79.67	89.34
Entertainment	96.46	94.80	97.33	69.91	68.22	84.39
Politics	93.33	88.88	93.33	94.16	89.06	94.21
Sport	93.00	90.74	95.14	69.23	69.01	81.66
Tech	96.46	94.80	97.33	69.91	68.22	81.11

estimated. As per experimentation, Naïve Bayes achieves 96.46% accuracy for Entertainment category and lowest accuracy for Business category, i.e., 92.63%. Similarly, with the Support Vector Machine 94.16% accuracy is achieved for Politics Category and Lowest Accuracy for Sport Category, i.e., 69.01%.

6 Conclusions and Future Scope

The work has been accomplished to find the polarity of the news articles. The result shows the category wise document polarity. The work is based on the Dictionary based approach with machine learning techniques. From above experimentation, it can be said that Naïve Bayes gives better results than Support Vector Machine as shown in Table 2. Future work will focus on other Classification techniques on other data related to online news articles and web blogs.

References

1. Roebuck, K.: Sentiment Analysis: High-Impact Strategies What You Need to Now: Definitions, Adoptions, Impact, Benefits, Maturity. Vendors, Emereo Publishing, 05 Nov 2012
2. Pooja, P., Sharvari, G.: A survey of sentiment classification techniques used for indian regional languages. Int. J. Comput. Sci. Appl. **5**(2) April 2015
3. Bo, P., Lillian, L., Shivakumar, V.: Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86 (2002)
4. Mohammad, S., Dorr, B., Dunne, C.: Generating high-coverage semantic orientation Lexicons from overly marked words and a thesaurus. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 599–608 (2009)
5. Turney, P.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the Association for Computational Linguistics, pp. 417–424, Philadelphia (2002)
6. Shoukry, A.: Collaboration Technologies and Systems (CTS). In: International Conference technologies and Systems, 21–25 May, pp. 546–550 (2012)
7. Alexandra, B., Ralf, S.: Rethinking Sentiment Analysis in the News, Theory to Practice and back||, European Commission, Joint Research Centre, Department of Software and Computing Systems, University of Alicante, WOMSA, pp. 1–12 (2009)
8. Ding, X., Liu, B., Yu, P.: A holistic lexicon-based approach to opinion mining. In: Proceedings of the International Conference on Web Search and Web Data Mining, pp. 231–240. ACM (2008)
9. Melville, P., Gryc, W., Lawrence, R.: Sentiment analysis of blogs by combining lexical knowledge with text classification. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 1275–1284 (2009)
10. Emma, H., Xiaohui L., Yong S.: The role of text pre-processing in sentiment analysis. Procedia Comput. Sci. Elsevier, **17**, 26–32 (2013) [14] Tetlock, P., Saar-Tsechansky, M., Macskassy, S.: More than words: quantifying language to measure firms fundamentals. J. Financ. **63**(3), 1437–1467 (2008)
11. Bing, L.: Sentiment Analysis and Opinion Mining, Apr 22 (2012)
12. Melville, P., Gryc, W., Lawrence, R.: Sentiment analysis of blogs by combining lexical knowledge with text classification. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1275–1284. ACM (2009)
13. Jagdale, R.S., Shirsat, V.S., Deshmukh, S.N.: Sentiment analysis of events from twitter using open source tool. Int. J. Comput. Sci. Mob. Comput. **5**(4), pp. 475–485 (2016)
14. Ye, Q., Zhang, Z., Law, R.: Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Syst. Appl. **36**, 6527–6535 (2009)
15. Bhumika, M., Jadav, V., Vaghela, B.: Sentiment analysis using support vector machine based on feature selection and semantic analysis. Int. J. Comput. Appl. **146**(13) (2016)
16. BholaneSavita, D., Deipali, G.: Sentiment analysis on twitter data using support vector machine. Int. J. Comput. Sci. Trends Technol. **4**(3) (2016)

Multi-constraint QoS Disjoint Multipath Routing in SDN

Manan Doshi, Aayush Kamdar and Krishna Kansara

Abstract Efficient path computation that sustains varying quality of service requirements is a key networking concern. Even though modern networks turned to multipath routing schemes as a first step in this path, existing solutions still resulted in sub-flows being directed to the same paths. Moreover, maintaining the quality of service subject to multiple criterions while selecting the paths and handling connection requests dynamically has proved to be challenging tasks. Addressing all these issues requires a centralized, real-time- and fine-grained control of the network facilitated by Software Defined Networks (SDN) that have emerged as a revolutionary networking paradigm. In this paper, we deal with the former issue by computing *k*-max min disjoint paths and for the latter we use an analytic hierarchy process. The proposed solution combines the two approaches for deployment in an SDN environment.

Keywords SDN · QoS routing · Analytic hierarchy process · Disjoint multipath routing · Multi-constraints

1 Introduction

Today's network research is directed towards achieving a single shared physical network supporting different heterogeneous applications with distinct traffic characteristics and QoS requirements. QoS routing considers key networking parameters such as bandwidth, cost, delay, energy consumption, etc., in the process of path computation [1]. Achieving network service quality that adapts to varying user requirements is a primary networking concerns [2, 3].

M. Doshi (✉) · A. Kamdar · K. Kansara
Shri Bhagubhai Mafatlal Polytechnic, Mumbai, India
e-mail: manandoshi1607@gmail.com

A. Kamdar
e-mail: ashbkamdar@gmail.com

K. Kansara
e-mail: kansarakrishna@rediffmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_40

Compared to traditional routing schemes, multipath routing reduces congestion in the network by shifting traffic to alternate paths thereby improving network resource utilization. This load equalization is considered to be effective in the improvement of network reliability and reducing energy consumption. A common solution to split flows across multiple paths is ECMP protocol. However, sub flows may be directed to the same paths, even with ECMP enabled [4, 5]. This problem is addressed by obtaining multiple paths with a certain level of disjoint-ness. Disjoint multipath routing offers several added advantages such as increased reliability, resilience to failure, reduced congestion and higher scalability as packets are forwarded through distinct paths [6].

All these requirements call for higher level management of the network for optimization of user-centric and programmable paths. This administration needs to adopt a software-oriented approach [6–8]. As a new and emerging network technology, software-defined networks provide a centralized control of the network by separation of data and control planes. The programmable control plane creates a dynamic, open and controllable network environment. Thus, it serves as the basis for dynamic path planning subject to multiple constraints that meet the QoS requirements. This makes load balancing of the core network more efficient.

The structure of this paper is as follows: Sect. 2 gives an overview of related work and our contributions. Section 3 describes the analytic hierarchy process for meeting QoS requirements. In Sect. 4 we explain the procedure for obtaining k-max min disjoint QoS paths. In Sect. 5 we present the benefits of SDN integration and finally Sect. 6 concludes the paper.

2 Literature Review

Achieving a common solution that meets QoS demands and as well provides efficient path optimization has become the focus of current research accelerated by the growth of SDN. Many solutions have been proposed considering different aspects of this problem. The work in [1] addresses two main issues associated with QoS routing in end to end communication, i.e., which links to develop to meet certain connectivity requirements and selection of paths that meet QoS demands. In [3], the authors use analytic hierarchy process to obtain a single value for multiple QoS requirements and then apply a heuristic procedure for obtaining optimized communication paths and backup paths. Also, solutions have been proposed to solve the QoS problem by using variations of routing metrics (as in [9]). The work in [9] focuses on isotonicity, a key property of algebra.

A multipath solution has been discussed in [6] where they compute k-maximally disjoint paths instead of k-max min disjoint paths. Also, authors in [10], aim to improve the throughput in shared bottlenecks by forwarding sub-flows from a same MPTCP connection through multiple paths. Suurballe [11] and Bhandari [12] conducted studies on the basic problems of link disjoint. However, they do not consider the QoS requirements. Disjoint path computation is NP-hard for the min-max problem [13, 14].

Also previous works have shown the benefits of SDN in improving network performance. Sonkoly et al. [15] use a test bed to show improved MPTCP connections by choosing the best path for the first sub flow in an OpenFlow network.

Our contributions can be summarized as follows:

(a) We target two NP-hard problems—multi-constrained optimal routing problem and disjoint path selection.
(b) Combining with the analytic hierarchy process AHP to obtain link weights that meet varying QoS demands.
(c) Convert the multi-criterion problem to obtain a single link value thereby reducing complexity.
(d) Then use a heuristic algorithm for obtaining k-max min disjoint QoS paths.

The concepts of AHP and k-max min paths and many terminologies have been borrowed from previous works [3, 10].

3 Analytic Hierarchy Process

We address the QoS problem by using the analytic hierarchy process, a decision analysis method. The network service quality requirements may be subject to multiple constraints that vary from person to person. AHP helps users assign accurate weights based on their focus on different criteria such as bandwidth, delay, cost, energy consumption, etc. The first step would be construction of hierarchies according to the requirements. Next is to perform a pair wise comparison of factors and weigh them on a scale of 1–9. We assume this scale as it is widely used however actual range may differ. Based on these comparisons we prepare a matrix A as illustrated in Eq. (1) that records the weights or ratings for N factors.

$$A = \begin{bmatrix} a_{11} & L & a_{1N} \\ M & 0 & M \\ a_{N1} & L & a_{NN} \end{bmatrix} \tag{1}$$

Here, $a_{ii} = 1$ and $a_{ij} = 1/a_{ji}$. We perform consistency checking using the following formulas. Our obtained values have passed the consistency checking when CR is less than 0.1 or else we need to reconsider the process.

$$\text{C.I.} = \frac{\lambda_{\max} - n}{n - 1} \tag{2}$$

$$\text{C.R.} = \frac{(a_1)(CI_1) + (a_2)(CI_2) + \cdots + (a_m)(CI_m)}{(a_1)(RI_1) + (a_2)(RI_2) + \cdots + (a_m)(RI_m)} \tag{3}$$

For our example we consider three criteria—bandwidth, delay, and energy consumption denoted as (α, β, ϒ) respectively. To reduce the complexity of processing multiple values we convert them into a single value called Multi-Criterion Cost MCC. We assign it as the link weight for path computation and are given as:

$$\text{MCC} = \text{b(e)} = \frac{\alpha * b(e)}{B} + \frac{\beta * d(e)}{D} + \frac{\gamma * g(e)}{G} \tag{4}$$

Here, b(e), d(e) and g(e) denote the remaining bandwidth, delay and energy consumption of the corresponding edge e. B, D and G are pre-defined constraints for the three factors respectively. In this way we assign new multi-criterion costs to links using AHP that meet QoS requirements of different users. The next phase involves computation of k-max min QoS disjoint paths.

4 *K*-Max Min Disjoint QoS Paths

We use a two-step algorithm for path computation. At first, a set of candidate paths are obtained between a source and destination pair using modified Dijkstra's algorithm. The next step is to use a greedy technique to select the *k*-max min QoS disjoint paths from the candidates of Step 1. The network is represented as a weighted graph *G (V, E)* where, *V* is a set of vertices (SDN Switches) and *E* is a set of edges (Switch links), and *b (u, v)* is the Multi-Criterion Cost (delay and bandwidth) for each edge *(u, v)* ϵ *E* allocated using AHP. Let *s* ϵ *V* and *t* ϵ *V* be the source and destination nodes respectively. Assuming there is a path from *s* to *t*, the minimum remaining MCC (Multi-Criterion Cost) of an edge along the path is the *Bottleneck Constraint* of that path. MBC Maximum Bottleneck Constraint is the maximum among all bottleneck constraints of the multiple paths between node *s* and node *v*. MHC indicates the minimum hop count of the shortest path from *s* to *v*.

In Fig. 1 each node v has a node ID indicated by the first alphabet and the next two numbers indicate v.MBC and v.MHC respectively. Initially for root node s, s.MBC = 0 and s.MHC = ∞, and for all other nodes, v.MBC = ∞ and v.MHC = 0. A node may belong to either of the following—visited, unvisited or marked.

Starting with root node *s*, as *s*.MHC <= *a*.MHC, we execute one-way relaxation of edge *(s, a)* thereby updating *a*.MBC and *a*.MHC to 8 and 1, respectively. Also, we record the parent node *s* and the bottleneck constraint 8 of the path from *s* to *a* as *a* (*s*: 10). Similarly, nodes *b* and *c* are visited and their corresponding MHC and MBC are updated. As all the neighbors of *s* are processed, we change the status of *s* as marked. Next, we select the neighbor of *s* with maximum bottleneck constraint, i.e., *a*. We continue with this process and the Fig. 2 shows the status of the nodes after *a*, *b,* and *d* relax their outgoing edges.

In Fig. 3, we illustrate two-way relaxation operation for nodes *c* and *e* as *e*.MHC > *c*.MHC. This results in edges *(e, c)* and *(c, e)* being relaxed simultaneously.

Fig. 1 SDN network with eight switches

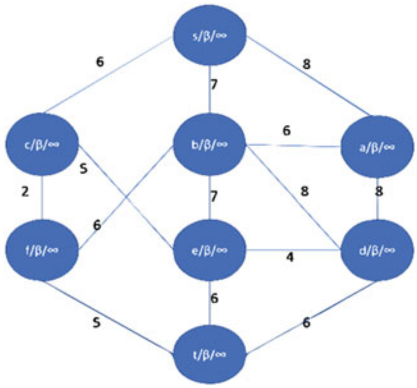


Fig. 2 Nodes *a*, *b* and *d* relaxes their outgoing edges

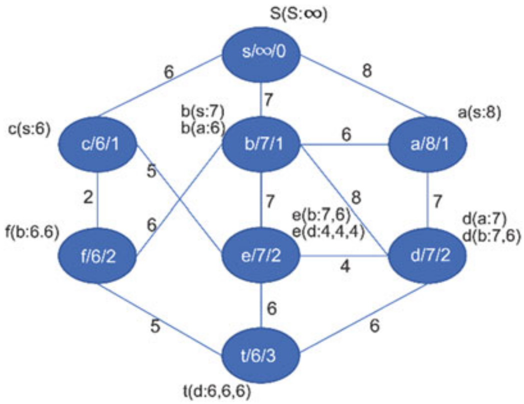
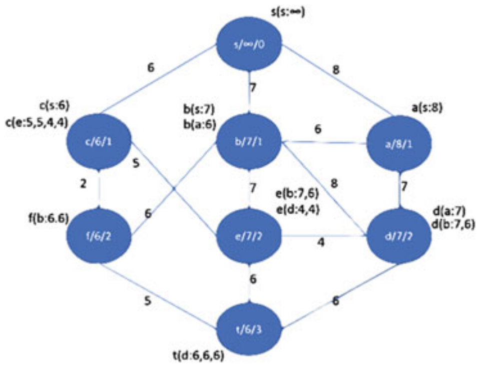


Fig. 3 Two-way relaxation of *(e, c)* and *(c, e)*



While relaxing edge *(e, c)* only the four paths with largest bottleneck constraints are selected. The selection of four paths only is an assumption we make for our algorithm. The final result of Step 1 is as show in Fig. 4.

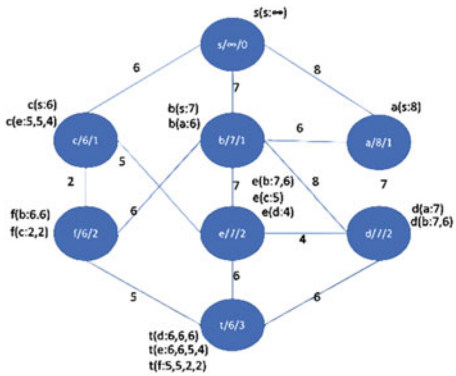


Fig. 4 Final result of Step 1

Algorithm 1: Obtain a set of candidate paths

1: **procedure** MAIN(s, t):
2: **for** all neighbours of u: v_i ($1 <= i <= n$) in increasing order of MHC **do**
3: **repeat** steps 8: to 20:
4: **if** multiple nodes with same MHC **then**
5: select randomly
6: **end if**
7: **end for**
8: **if** u.MHC <= v.MHC **then**
9: **call** One-way relaxation (u, v)
10: **else**
11: **call** Two-way relaxation (u, v)
12: **end if**
13: **set** node x with maximum MBC (then least MHC) from v_1 to v_i ($1 <= i <= n$)
14: **add** u to *Marked*
15: $u \leftarrow x$
16: **if** all nodes in *Marked* **then**
17: **stop**
18: **else**
19: **goto** step 2
20: **end if**
21: **end procedure**
22: **function** One-way relaxation(u, v):
23: **if** v.MBC < min *(u.MBC, b (u, v))* **then**
24: **set** v.MBC = min *(u.MBC, b (u, v))*
25: **end if**
26: **if** v.MHC > u.MHC + 1 **then**
27: **set** v.MHC = u.MHC + 1
28: **end if**
29: **store** parent node as {u: (Bottleneck Bandwidth of paths from s to u)}
30: **return**
31: **function** Two-way relaxation (u, v):
32: **call** One-way relaxation (u, v)
33: **call** One-way relaxation (u, v)
34: **return**

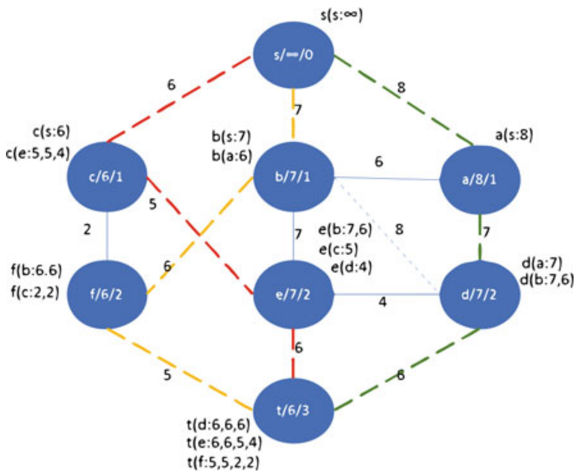
Algorithm 2: Obtain k-max min QoS disjoint paths from candidates

Let x be the number of candidate paths obtained at destination T at the end of algorithm 1, k be the number of outgoing edges of the root node S, *max* be the maximum of the MBC of a set of disjoint paths and *P_temp* holds the set of disjoint paths.

```
procedure MAIN(s, t):
for i, j in range(1,x) do
        if cp_i ≠ cp_j then      /*if cp_i and cp_j are disjoint paths
            T_disj [i][j] = True
        end if
end for
for i in range(1,x) do
        p_1 = cp_i               /* Candidate path: cp_i
        for j in range(1,k) do
            Obtain a path p_j disjoint from paths p_1 to p_(j-1)
            check (j-1) columns:
            T_disj [1][p_1] to T_disj [x][p_1] and;
            T_disj [1][p_(j-1)] to T_disj [x][p_(j-1)]
            add p_j to P_temp
            if path not found then break
            end if
            Obtain max of paths p_1 to p_k
            if new max > current value then
                update max and P_temp to new values
            else
                retain old values of max and P_temp
            end if
        end for
    end for
    Obtain k-max min disjoint QoS paths in P_temp
end procedure
```

We now use the greedy technique to select k disjoint QoS paths from the set of candidate paths and try to maximize the minimum bottleneck constraint path of the k disjoint paths. Let x be the number of candidate paths $cp_1, cp_2, \ldots, cp_x$ obtained at the end of Step 1 to reach destination node t from the source node s. The candidate paths are arranged in the decreasing order of bottleneck constraints. *T_disj [] []* is a two-dimensional array that stores the disjoint path relations among the candidates. This step uses x iterations where we obtain x sets of k disjoint paths, stored in *P_temp []*, and their minimum bottleneck constraints. The maximum of the minimum bottleneck constraints is recorded and the corresponding set of k-max min QoS disjoint paths are obtained in *P_temp []*. The final result of Step 2 is as illustrated in Fig. 5 analyzing the time complexity of the algorithm, Step 1 takes $O(|E| \log |V| + |V| \log |V|)$ while Step 2 runs in $O(|V|^2)$.

Fig. 5 Final result of Step 2



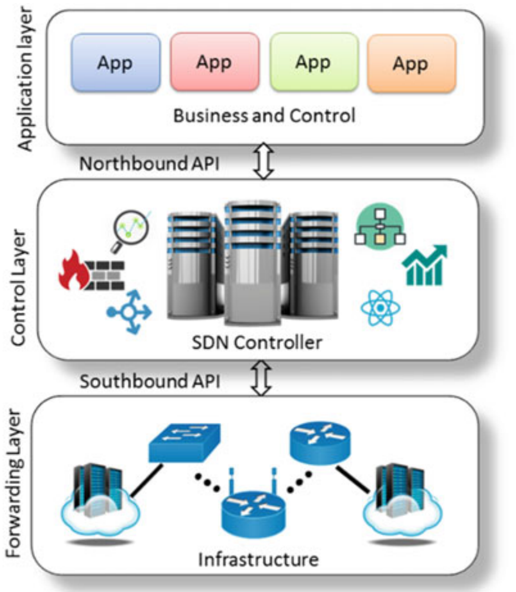
5 Benefits of SDN

The Open Networking Foundation (ONF) defines SDN as an emerging network architecture where the network is directly programmable and where the control and forwarding planes are decoupled [16, 17]. The network intelligence and states are handled by controllers that globally regulate network states. The separation allows data plane devices to be designed as simple, unintelligent devices (SDN switches) that are solely left with the task of forwarding packets based on decisions taken by the controllers [18]. A set of north-bound application programming interfaces (APIs) serve as the Application-Controller Plane Interface (A-CPI) while the Data-Controller Plane Interface (D-CPI) is handled by south-bound APIs such as OpenFlow protocol. The OF protocol allows the logically centralized controller to dynamically modify the forwarding table of routers and switches. OF uses match-action abstraction to aggregate flows across the data plane [19–21]. The logical view of SDN is illustrated in Fig. 6.

Our intuition for better performance of the proposed system is based on the fact that the two techniques we use have been investigated in previous works [3, 10]. AHP and k-max min disjoint path computation were realized as promising solutions that improved overall network performance. We combine these techniques for implementation in SDN so that the benefits of the promising technology can further enhance performance.

The proposed algorithms can be designed in Mininet. The Mininet creates a realistic virtual network, running the real kernel, switches and application code, on a single machine (VM, cloud or native). This makes us customize our SDN environment. OpenFlow can be used to manage all flows while decision-making and setting of rules would be handled by the SDN controller. The simulations should be preferably carried out in real SDN topologies to obtain accurate results. Initially, generate some random traffic in the network. Next, for a random pair, monitor the end-to-end

Fig. 6 SDN architecture



communication based on various QoS constraints such as bottleneck bandwidth, jitter, delay, throughput etc. using IPerf for a stipulated time period. Also traffic flow can be analyzed real time using tools such Wireshark. VeriFlow can be used to distinguish between elephant and mice flows for better traffic handling. Finally analyze the collected statistics and verify whether the results satisfy the QoS requirements [10].

6 Conclusion

With SDN accelerating the innovation and evolution of modern networks, development of a highly scalable and intelligent TE system that sustains varying QoS requirements and maintains efficiency of path selection can be envisioned. Therefore, in this paper we propose a two-phase procedure for achieving these objectives. The first phase uses the analytic hierarchy process to capture the varying QoS requirements and reduce it to a new cost function that is used to assign link weights. This reduces the complexity of processing and improves quality control of the network. The second phase obtains *k*-max min disjoint paths that makes the load balancing of the core network efficient, reduces congestion and improves reliability.

In the future, we focus on implementing the proposed solution in an SDN environment. SDN has a flexible and open architecture that allows dynamic regulation of network behavior. It provides software oriented control of the network and facilitates centralized path routing. As traffic patterns change, dynamic computation of paths can be done to meet QoS needs. With these benefits, the efficiency of our proposal can be embodied to the maximum degree.

References

1. Hung, M.-H., Wang, C.-H., He, Y.: A real-time routing algorithm for end-to-end communication networks with QoS requirements. In: 3rd International Conference on Computing Measurement Control and Sensor Network (2016). https://doi.org/10.1109/cmcsn.2016.44
2. Orda, A., Sprintson, A.: Author. Efficient algorithm for computing disjoint QoS paths. In: Proceedings of IEEE INFOCOM 2004, vol. 1, pp. 727–738. IEEE Press (2004)
3. Pan, Q., Zheng, X.: MULTI-path sdn route selection subject to multi-constraints. In: 3rd International Conference on Cyberspace Technology (2015). https://doi.org/10.1049/cp.2015.0818
4. Sandri, M., Silva, A., Rocha, L.A., Verdi, F.L.: On the benefits of using multipath TCP and Openflow in shared bottlenecks. In: 29th International Conference on Advanced Information Networking and Applications. IEEE (2015)
5. Chiesa, M., Kindler, G., Schapira, M.: Traffic engineering with equal-cost-multipath: an algorithmic perspective. In: INFOCOM, 2014 Proceedings IEEE, pp. 1590–1598, Apr 2014
6. Abe, J.O., Mantar, H.A., Yayimli, A.G.: k-Maximally disjoint path routing algorithms for SDN. In: International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (2015)
7. Campista, M.E.M., Rubinstein, M.G., Moraes, I.M., Costa, L.H.M.K., Duarte, O.C.M.B.: Challenges and research directions for the future internetworking. Commun. Surv. Tutor. IEEE **16**(2),1050– 1079, Second 2014
8. Akyildiz, I.F., Lee, A., Wang, P., Luo, M., Chou, W.: A roadmap for traffic engineering in SDN-OpenFlow networks. Comput. Netw. **7**(**1**), pp. 1–30, 4 Oct 2014. ISSN 1389-1286
9. Geng, H., Shi, X., Yin, X., Wang, Z., Yin, S.: Algebra and algorithms for multipath QoS routing in link state networks. J. Commun. Netw. **19**(2) (2017)
10. Sheu, J.-P., Liu, L.-W., Jagadeesha, R.B., Chang, Y.-C.: An efficient multipath routing algorithm for multipath TCP in software-defined networks. In: European Conference on Networks and Communications (EuCNC) (2016)
11. Suurballe, J.W.: Disjoint paths in a network. Networks **4**(2), 125–145 (1974)
12. Bhandari, R.: Survivable networks: algorithms for diverse routing. In: Springer Science & Business Media (1999)
13. Guo, Longkun, Shen, Hong: On finding min-min disjoint paths. Algorithmica **66**(3), 641–653 (2013)
14. Li, C.-L., McCormic, S.T., Simchi-Levi, D.: The complexity of finding two disjoint paths with min-max objective function. Discret. Appl. Math. **26**(1), 105–115 (1990)
15. Sonkoly, B., Nemeth, F., Csikor, L., Gulyas, L., Gulyas, A.: SDN based testbeds for evaluating and promoting multipath TCP. In: 2014 IEEE International Conference on Communications (ICC), pp. 3044–3050, June 2014
16. Mendiola, A., Astorga, J., Jacob, E., Higuero, M.: A survey on the contributions of Software-Defined Networking to Traffic Engineering. In: IEEE Communications Surveys & Tutorials, vol. 19, Issue 2, Second quarter 2017. https://doi.org/10.1109/comst.2016.2633579
17. Open Networking Foundation, Software-Defined Networking: The new norm for networks, Technical. report, 2012, white paper
18. Michel, O., Keller, E.: SDN in wide-area networks: a survey. In: 4th International Conference on Software Defined Systems (SDS) (2017). https://doi.org/10.1109/sds.2017.7939138
19. Akyildiz, I.F., Lee, A., Wang, P., Luo, M., Chou, W.: Research challenges for traffic engineering in software defined networks. In: IEEE Network, vol. 30, Issue 3, May-June 2016. https://doi.org/10.1109/mnet.2016.7474344

20. OpenFlow Switch Specification v1.0-v1.4. www.opennetworking.org/sdn-resources/onf-specifications
21. Akyildiz, F., et al.: A roadmap for traffic engineering in software defined networks. Comput. Netw. **71**, 1–30 (2014)

Performance Analysis of Trust-Based Routing Protocol for MANET

Archana Mandhare and Sujata Kadam

Abstract Mobile Ad hoc network (MANET) is a self-motivated network. Nodes are freely moved anywhere inside the network. They can enter and depart the network at any instance. Due to loss of infrastructure in the network frequent link failures occurred. So the special classes of routing protocols are taken into account for reducing the link failures. Categories of protocols are Reactive, Proactive, and Hybrid. This reactive protocol reduces the routing overhead via sending the routing packets on every occasion there is want of communication. This paper mainly focused on the reactive protocols such as AODV, DSR, and AOMDV. Proposed work utilizes trust concept for finding the reliable route. In this work, we implemented Secure Routing Protocol which establishes a secure path between the nodes, totally based on the node trustworthiness. Based on the nodes past experience we calculate the nodes present trust value and finds the better route for data transmission. Proposed method improves the packet delivery ratio and throughput.

Keywords Ad hoc network · Trusted node · Routing protocols Packet dropping · Reliable path

1 Introduction

MANET does not have any central control. MANETs are also called as MWN [1]. In MANET any node can be act as a host and router. Figure 1 shows that number of portable devices are connected without any centralized structure. From Fig. 1 we can see that cell phones, laptops, PDAs acting as a wireless medium for communication [1].

A. Mandhare (✉) · S. Kadam
Department of Electronics & Telecommunication Engineering, Ramrao Adik Institute of Technology, Nerul, Navi Mumbai 400706, India
e-mail: mandhare.ap@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_41

Fig. 1 Architecture of MANET [1]

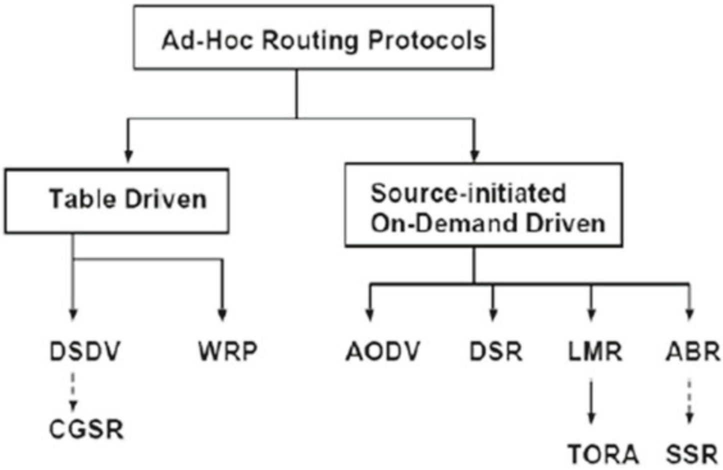


Fig. 2 Categories of routing protocols [1]

2 Routing in MANETs

Routing is the strategy used to exchange a data from the source to destination. It allows the messages to go from one node (PC) to each other driving node to reach up to the destination [2]. Different types of protocol in MANET are as shown in Fig. 2. It relies on various factors like topology, choice of routes, the start of route request [3].

2.1 Proactive Routing Protocol

This kind of protocols exposed routes before communication. Every node keeps a table in which routes are modified on every time whenever any trade occurs. The routing tables are updated periodically each time when the network topology changes. This category of protocol is not applicable for large networks [4].

2.2 *Reactive Routing Protocol*

These protocols are additionally referred to as on request protocols. Source node initiates direction discovery on call for foundation [1]. In the route finding method starting place node starts route discovery when there is a requirement of a route. These protocols lower the routing overhead but on the fee of expanded put off inside the network [4]. This sort of protocols is intended to limit routing overhead. This paper discussed about three on-demand ad hoc routing protocols AODV, AOMDV, and DSR as follows [5].

2.2.1 Ad hoc on Demand Distance Vector Routing (AODV)

This protocol minimizes the broadcasts by finding routes totally basis on need. In this route is discovered when there is applicable route entry in the routing table [5]. A routing table finds out whether the path is already available to that destination or not. If there is a valid route that route is used for data transmission [6].

Figure 3a shows that the broadcasts RREQ packet in the network. Transmission starts from source node S in the network. Figure 3b exhibits the example in which we see that nodes D sending a RREP packet to the C and F nodes. Both nodes having a path to S node, i.e., source node.

2.2.2 Ad Hoc on Demand Multipath Distance Vector Routing (AOMDV)

This protocol manages a couple of loop-free paths. The advantage of the utilization of AOMDV is that it allows intermediate nodes to respond to RREQs packets. But, AOMDV has larger message overhead [5].

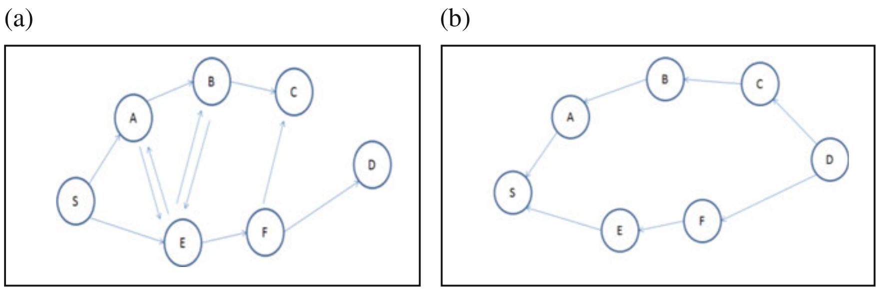


Fig. 3 **a** Route request broadcasts and **b** route reply propagation

2.2.3 Dynamic Source Routing (DSR)

This sort of routing contains a route cache for every node. If the node having route entry already in the routing table it discards the second route request it does not rebroadcast it again. This protocol has less message overhead [5].

3 Proposed Methodology

Proposed node architecture consists of six blocks as shown in Fig. 4. Each node contains six modules they are as follows:

- *Process*: A process is an application-layer entity (running program). It contains all the routing related information.
- *Buffer*: A buffer is a set of memory locations. It is a region of a physical memory storage used to temporarily store data.
- *Encryption/Decryption*: We have used Open SSL Library for the purpose of encryption and decryption of the data. This library is in-built in Linux. For Encryption and decryption purpose we have used Public Key Algorithm (RSA) [10].
- *Transmitter and Receiver*: In the first stage of data transmission, RREQ packet is transmitted to the intermediate nodes in the path.
- *Trusted Party* (*Update Trust Value*): The trusted party works as the centralized system for all the nodes. It up-dates trust values of every node [7]. For communication purpose node has associated with the trusted party. Updated trust values are used for route establishment. The nodes behave well in past are considered in routing [7].

3.1 Trust Management

Trust assumes an exceptionally conspicuous part in MANET. Trust value shows the genuineness of the particular node. Trust control schemes are devised to find out

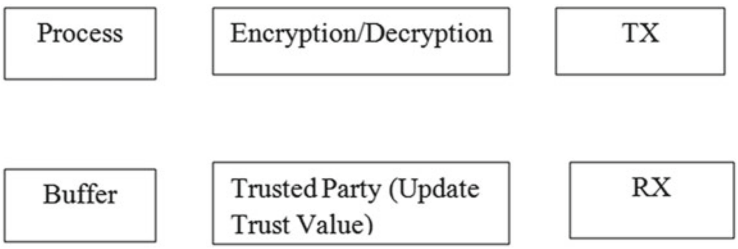


Fig. 4 Proposed node architecture

misbehaving nodes. Trust administration incorporates Trust creation, Trust update, and Trust revocation. This work uses trust management system [8].

3.2 Trust Value Calculation

Trust values are calculated based on following assumptions [7]:

- $\Gamma_k{}^{(1)}$ = (Number of packets that are relayed in last w sessions)/(Total no of incoming packets in last w sessions) (1)
- $\Gamma_k{}^{(2)} = 1 -$ ((Number of sessions broken by N_k in the last w sessions)/w)) (2)
- $\Gamma_k{}^{(3)}$ = (Number of session that N_k relayed at least δ packets/w) (3)
- $\Gamma_k{}^{(4)}$ = (Number of session that N_k participated in the period w/M) (4)

N_k: Trustworthiness of node.
Γ_k: N dimension vector of trust values [$\Gamma_k{}^{(1)}$, $\Gamma_k{}^{(2)}$ … $\Gamma_k{}^{(n)}$].
$\Gamma_k{}^{(i)}$: Assigned any real value between [0, 1].
w: Number of sessions.

Packet format of trust-based routing protocol contains the source id (IDS), Destination id (IDD), Next hop, Encrypted key (AES) and Decrypted key (DES), time stamp (ts), trust value (Tr), and energy requirements (Er).

With the help of above equations, we assigned trust values to each node based on node past performance. The node drops the packets due to load on the node, busy channel in that case node trust value will be low. Nodes which are participating in the network more time having higher trust value. In other reactive protocol, they are not using trust technique for finding the good nodes for data transmission.

3.3 Metrics for Performance Evaluation

Performance is examined by utilizing the accompanying measurements. Performance metrics are as follows:

- Packet Delivery Ratio.
- End-To-End delay.
- Throughput.
- Routing Overhead.

4 Results and Discussion

Performance analysis is done with the help of Network Simulator (NS-2) software [9, 10]. Every simulation session we set the source and destination node. Routing

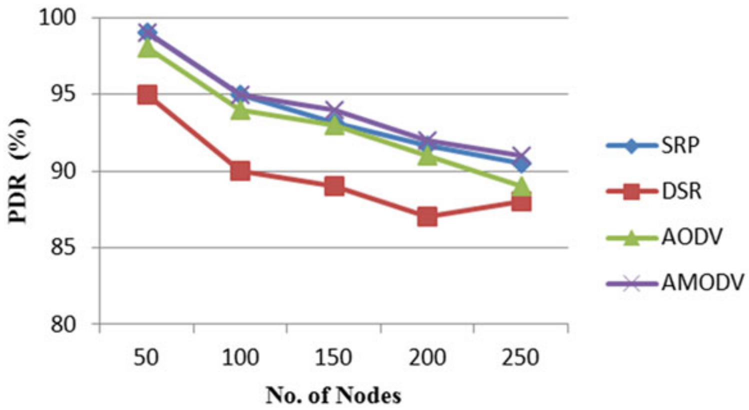


Fig. 5 Packet delivery ratio

algorithm finds out the trusted intermediate node for data transmission. To validate the performance of the proposed technique, we differentiate this with the existing different routing protocols such as DSR, AODV, AOMDV.

4.1 Packet Delivery Ratio

From the above excel graph we observed that we get higher packet delivery for SRP and AOMDV protocol. Fewer packets are delivered in case of DSR (Fig. 5).

4.2 End-to-End Delay

From the below graph we can see that delay is somewhat high in all the protocols as contrasted to existing protocols. Presented work uses trusted node for data transfer. Delay is measured in milliseconds (Fig. 6).

4.3 Throughput

Figure 7 demonstrates that as the number of nodes increases throughput decreases. The graph shows that all protocols have some similar throughput. It is changing with small values. Throughput measured in kbps.

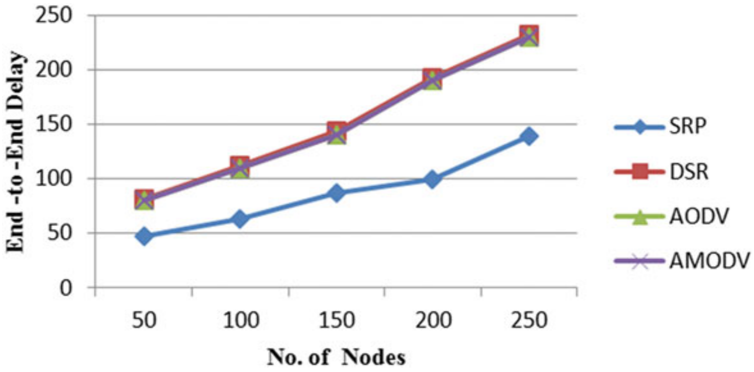


Fig. 6 End-to-end delay

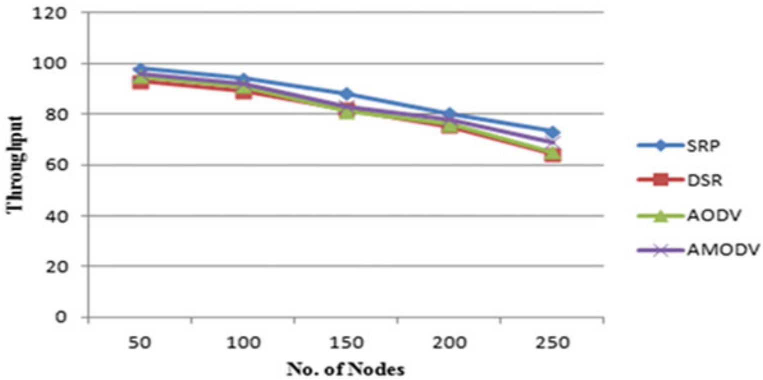


Fig. 7 Throughput

4.4 Routing Load

Figure 8 demonstrates that routing load for DSR protocol is improved as equated to other protocols. As reduce the number of routing packets in the network. The life time of network will automatically increased.

Table 1 we observed that comparative results for four network parameters. Above results are for number of 50 nodes simulation. We get better results in PDR and Throughput for proposed protocol as compared to existing. As we have used trust concept for routing which increases the packet delivery of nodes. But the two parameters decrease the performance of network as compared to the existing method.

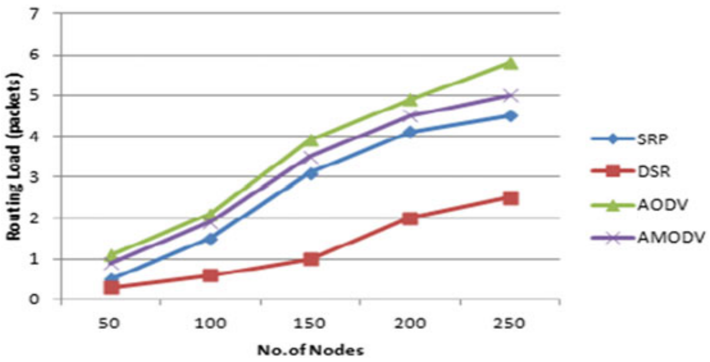


Fig. 8 Routing load

Table 1 Comparative analysis of proposed and existing protocol for 50 nodes

Network parameters	No. of nodes	SRP	DSR	AODV	AOMDV
Packet delivery ratio (%)	50	99	95	98	99
End to end delay (ms)	50	48	81	80	80
Throughput (Kbps)	50	98	93	95	96
Routing load (Routing packets)	50	0.5	0.3	1.1	0.9

5 Conclusion

In this paper, we have presented secure routing based on the trustworthiness of the node, which has been designed for MANETs to find the reliable route for data transmission. We compared the performance of network on various parameters such as pdr, throughput, delay, routing overhead with the existing DSR, AODV, AOMDV protocols. We get better performance in different network parameters such as good packet delivery in proposed method than the existing methods. Delay is less as differentiated to other protocols.

6 Future Scope

There are some open issues still present in presented work; we observed from the simulation results that, if an already established path comes again and again in the network then the nodes in that path will be overloaded. So in that case to reduce the overload of the node, we can consider energy and trust of the node. If we consider these parameters ultimately network life will increase.

References

1. Dhenakaran, F., Parvathavarthini, S.: An overview of routing protocols in mobile ad-hoc network. Comput. Sci. Softw. Eng. **3**(2) (2013)
2. Liu, F., Deng, S., Balakrishnan, T.: An acknowledgement-based approach for the detection of routing misbehavior in MANETs. IEEE Trans. Mobile Comput. **6**(5), 536–550 (2007)
3. Marti, F., Giuli, S., Lai, T.: Mitigating routing misbehavior in mobile ad hoc networks. In: Proceedings. ACM A (2000). https://doi.org/10.1109/icrtit.2011.5972464
4. Singh, F., Singh, S.: MANET: a study of challenges and routing principles. Int. J. Adv. Innov. Res. **1** (2012)
5. Chadha, F., Joon, S., Sandeep, T.: Comparison of AODV, DSR and AOMDV routing protocols in MANETs. Int. J. Soft Comput. Eng. (IJSCE) **2**(3), J (2012)
6. Singh, F., Jain, S.: Secure AODV routing protocols based on concept of trust in MANET's. Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) **3**(12) (2014)
7. Mohamed, F., Xuedong, S., Shen T.: Secure and reliable routing protocols for heterogeneous multihop wireless networks. IEEE Trans. Parallel Distrib. Syst. **26**(4) (2015)
8. Shitole, F., Purohit, S., Gite, T.: A trust based model for self organized mobile ad-hoc network. Int. J. Emerg. Technol. Adv. Eng. **5**(12) (2015)
9. https://www.isi.edu/nsnam/ns/tutorial/
10. https://www.openssl.org/source

VANET-Based Distributed Platoon System

Vanshri Deshpande and Swati Kamthekar

Abstract Intelligent Transportation Systems (ITS) targets to streamline the vehicle operation, also assist driver with various safety, on board, and surrounding information. On highway grouping of vehicles forms a Platoon. Vehicular Ad hoc Network (VANET) enables Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I) communication that helps to increase the performance of platoon. Platoon helps to improve road capacity, more comfortable and safe driving, reduce in emissions. In previous year the most of the research is done on string stability analysis, latency, and data dissemination. The control of position, speed, and acceleration to the preceding vehicle in platoon is referred as longitudinal control. This work focus on the basic and mini platoon control design strategy. The design is proposed with Ad hoc demand distance Vector (AODV) routing protocol, used for application of platoon the new developed algorithm Distributed platoon system (DPTS) is proposed. The comparative study of strategy based on parameters like Packet delivery ratio (PDR), Throughput, Routing Overhead, End-to-end (e2e) delay. Traffic disturbance scenario for failure of vehicle in platoon is developed.

Keywords Vehicle platoon · Leader vehicle · AODV · DPTS · Platoon strategy
Route request (RREQ) · Route reply (RREP)

1 Introduction

In Automated Highway system (AHS) group of consecutive vehicles forms Platoon. Vehicle platoon consists of leader vehicle, tail vehicle and members of sys tem. Platoon control exchange the state of information like distance, velocity, and

V. Deshpande
Saraswati College of Engineering, Kharghar 410206, Maharashtra, India
e-mail: vnshr.d@gmail.com

S. Kamthekar (✉)
Pillai HOC College of Engineering and Technology, Raigad 410206, Maharashtra, India
e-mail: swati.kamthekar@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_42

acceleration with consecutive vehicles through wireless link. In platoon the vehicles operates close together with preceding vehicle, this increases road capacity as it can accommodate more vehicles on highway, it also reduce the energy consumption, the advance technologies helps to drive more safer and comfortable.

Platoon system implementation is described with Adaptive cruise control (ACC) and Co-operative adaptive cruise control (CACC). In ACC system vehicle ensures constant time gap with respect to immediate preceding vehicle, which leads to string stability. CACC system is ACC system enabled with wireless communication. CACC-based platoon system with analysis of string stability is proposed in [1]. In CACC system link is established to communicate leading vehicle information like acceleration and speed, with platoon id, source and destination address [2]. Communication techniques like Dedicated short range communication (DSRC), Wireless access in vehicular environment (WAVE) for platoon is proposed [3].

A Novel Disturbance adaptive platoon system with two adjacent platoons on same lane or on different lane with analysis of platoon dynamics like platoon size and intra platoon spacing is proposed [4]. A Markov based Platoon system is proposed in [5]. System is developed were sensors communicates with each other through a Markov-based random access protocol [6] the effect on stability of platoon due to delay is studied in [7] and Packet drop out in platoon is addressed in [8].

Vehicle Ad hoc network (VANET) enables vehicle to vehicle (V2V) and vehicle to Infrastructure (V2I) communication. In V2V the information is shared between two vehicles and in V2I the information is shared between Vehicle and Road side unit (RSU) [9].

A vehicle platoon model based on robots is developed and analysis is done for velocity stability based on the data received from leader and from member of platoon [10]. Control algorithm based on Common Quadratic Lyapunov Function (CQLF), is developed and evaluated for safe and stable operation of platoon for wide range [11]. Based on distributed control named as sliding mode control (SMC) is used to achieve the stability between members of platoon and string stability of entire platoon system as detailed in [12]. Various different types of reactive routing protocols like AODV and Destination Sequence Distance Vector (DSDV) are applicable in platoon management system. AODV gives better performance (PDR, routing overhead and throughput) as compare to DSDV [13].

2 Related Work

2.1 Basic Platoon Design Strategy

The design strategy defines the way of signal (i.e., speed, acceleration, distance, etc.) communication between vehicles of platoon. Basic platoon design strategy is shown in Fig. 1 also known as Leader following strategy, in which the information is communicated from leader to next immediate following vehicle and so on till the last

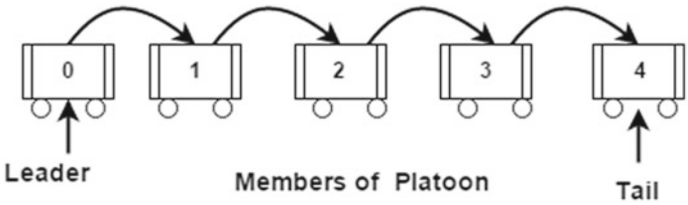


Fig. 1 Existing basic platoon design strategy

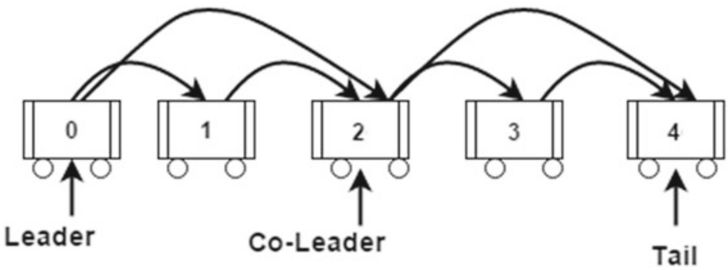


Fig. 2 Proposed mini platoon design strategy

vehicle of platoon. The Leader is first vehicle of platoon which generates the information like speed of vehicles, minimum distance to be maintained between vehicles, the direction of platoon, destination of platoon, in case due to communication failure if platoon breaks leader communicates all the information with Road side units (RSUs). The last vehicle of platoon is Tail vehicle which receives information from preceding vehicle and follows the same. In case of adjacent platoon system the tail vehicle communicates the received information to leader of adjacent platoon system.

2.2 Mini Platoon Design Strategy

In it the system consists of one leader and number of co-leaders. The leader is the first vehicle of platoon which creates the platoon and conveys information related to speed, distance, direction, etc., to following vehicles function is same as of leader of basic platoon strategy. The co-leader is one which acts as leader for immediate following vehicles and it communicates the same information received from leader till the last vehicle of platoon. A platoon may consist of number of co-leaders based on communication (WAVE, DSRC, IEEE 802.11, etc.) accomplish on vehicle. The vehicle which is last or at the boundary of coverage area of leader will act as co-leader and so on till the last vehicle of platoon. So, in mini strategy a large platoon is divided in to the group of vehicles forming a mini platoon with co-leaders as shown in Fig. 2.

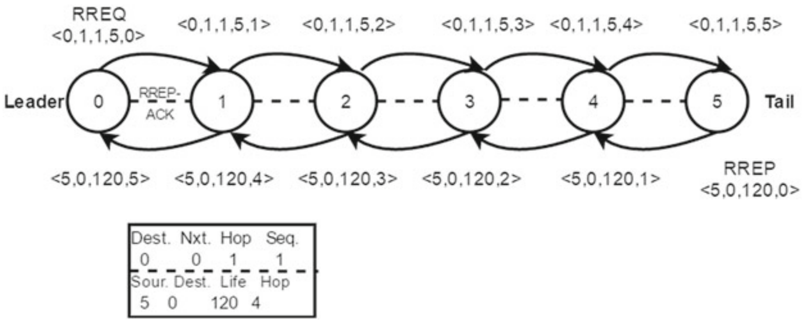


Fig. 3 DPTS routing technique for basic platoon design

3 Design and Simulation

3.1 DPTS Routing Technique for Basic Platoon Design

In this work, the Ad hoc demand distance vector routing protocol (AODV) has been applied for the design of platoon system named as DPTS. As shown in Fig. 3 the system consists of six nodes, leader, tail and four members of platoon. AODV has path and data packets. The path, i.e., route is developed first and then data packets are transferred. It makes use of various message packet formats like route request (RREQ), Rote reply (RREP), Route reply acknowledgement (RREP-ACK). Simulation window of the same as shown in Fig. 3.

Steps:

1. Leader generates RREQ packet with details as: <Originator, Seq. No., Source ID, Destination, Hop count>. Packet format is <0, 1, 1, 5, 0>.
2. Nodes in first hop, i.e., node 1 receives the packet with <0, 1, 1, 5, 1>. The Node 1 will broadcast the RREQ packet.
3. Node 1 will make routing table entry to reply back to node 0: <Destination, Next node, hop count, sequence number> as <0, 0, 1, 1>.
4. Now. Nodes in second hop will receive the RREQ packet send by node 1. The node 1 will receive reply from node 2 as <0, 1, 1, 5, 2>.
5. The same procedure will be repeated till node 5. It will generate RREP packet: <Source ID, Destination ID, Life time, Hop count> the packet will be <5, 0, 120, 0>.
6. Node 0 is destination node through following nodes it will receive packet with values as <5, 0, 120, 5>.
7. Node 1 will maintain the routing table as: <Source node, destination node, life time, hop count> with values as <5, 0, 120, 4>.

Fig. 4 Simulation window of basic platoon design

8. Node 0, i.e., leader will get the information that it is possible to reach destination in five hops, so it will find the shortest route, other than described route there will be multiple routes.
9. Node 0 will reply with RREP-ACK packet to node 1. Node 1 will reply RREP-ACK packet to node 2 and so on till the last node of the route.
10. The routing table will be generated for each and every node. In Fig. 4 routing table for node 1 is shown.

3.2 *DPTS Routing Technique for Basic Platoon Design*

As shown in Fig. 5 the system consists of nine nodes, leader, tail, Co-leaders and members of platoon system. The number of co-leader depends upon the path formed, i.e., number of co-leaders may differ with communication techniques (coverage area) and with position (distance between nodes) of nodes. The design of system is carried using AODV routing protocol which has path and data packets. Simulation window of the same is shown in Fig. 6.

Steps:

1. Leader generates RREQ packet with details as: <0, 1, 1, 9, 0>.
2. Nodes in first hop, i.e., node 1 and node 2 receives the packet with <0, 1, 1, 9, 1>. Now, both the Node 1 and 2 will broadcast the RREQ packet.

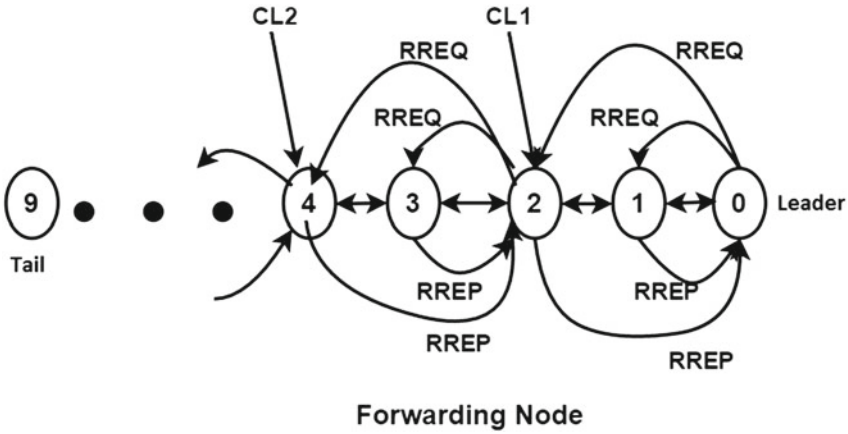


Fig. 5 DPTS routing technique for mini platoon

Fig. 6 Simulation window of mini platoon design

3. Again, Next hop nodes will receive the packet and broadcast the RREQ packet to its hop coverage nodes and so on till the destination node 9 is reached.
4. When RREQ received at node 9, it is a destination node. It will generate RREP packet: <9, 0, 120, 0>.
5. Node 0 has node 1 and 2 in its first hop so, it will receive reply from both the nodes. From node 1 it will receive RREP packet, e.g., <9, 0, 120, 9> and from node 2 it will receive RREP packet, e.g., <9, 0, 120, 8>. So, as reply received

from node 2 has less hop counts than Node 1, Node 0 will reply RREP-ACK to node 2.
6. Node 2 will now act as a co-leader and will broadcast the packet in next hop. Node 1 will receive all the information from leader (node 0), in case if node 2 fails the node 1 will act as a co-leader 1 (CL1).
7. Similarly, CL2, CL3 will be form till the destination. Co-leader is last vehicle or vehicle at the boundary of hop. Number of co-leaders depends upon the network.
8. The routing table will be created and maintained for each and every node.

4 Parameter Analysis

Analysis of System is carried based on parameters like end to end delay, Packet delivery ratio, Throughput, Routing overhead. The Tables 1 and 2 shows the comparison of basic and mini platoon design strategy. The Network Simulator 2 (NS-2) software is used for design and analysis of parameters.

Table 1 Simulation values of end to end delay, packet delivery ratio

No. of nodes	End to end delay (ms)		Packet delivery ratio	
	Basic platoon	Mini platoon	Basic platoon	Mini platoon
50	80	50	98	100
100	110	60	94	96
150	140	90	93	94
200	190	100	91	92
250	230	150	89	91.5

Table 2 Simulation values of throughput, routing overhead

No. of nodes	Throughput		Routing overhead	
	Basic platoon	Mini platoon	Basic platoon	Mini platoon
50	95	100	1.2	0.5
100	91	91	2.2	1.5
150	81	90	3.8	3.2
200	76	82	4.8	4.2
250	65	76	5.6	4.5

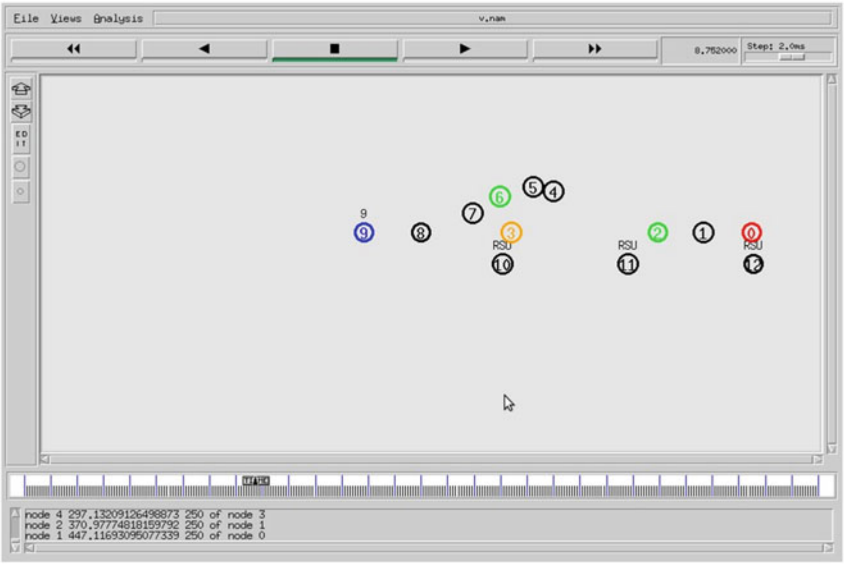


Fig. 7 Vehicle failure in platoon system scenario

5 Traffic Disturbance Scenario

5.1 Vehicle Failure Scenario

Scenario is developed with 10 vehicles and 3 RSU as shown in Fig. 7 RSU holds all the information of platoons in its coverage area. In platoon if one of the vehicle fails, leader communicate about the same to RSU. RSU communicate the information to following vehicles. So, with mini platoon strategy even if any vehicle in platoon fails the main platoon works without failure. The same scenario is developed in NS2 simulator.

6 Conclusion

In this work, two different platoon strategy basic design and mini platoon are designed. Comparative study of existing and proposed platoon design strategy based on various performance parameters like end-to-end delay, Packet delivery ratio, throughput, routing overhead is carried. More improved performance is achieved in Mini platoon design strategy. Also traffic disturbance scenario for failure of vehicle in platoon is developed.

The Ad hoc demand distance vector routing (AODV) protocol used for application of designed strategy, the new developed algorithm is called as DPTS (Distributed platoon system).

The proposed designed system makes it applicable to increase road capacity, provides comfortable driving for long routes and will also help to minimize the occurrence of accidents due to human error. In the future work the research on formation of platoon based on destination before the vehicle enters on highway and analysis in case of failure of leader vehicle.

References

1. Ploeg, J., Shukla, D.P., et al.: Controller synthesis for string stability of vehicle platoon. IEEE Trans. Intell. Transp. Syst. **15** (2014)
2. Ariffin, M.H.M., Rahman, M.A.A., Zamzuri, H.: Effect of leader information broadcasted throughout vehicle platoon in a constant spacing policy. In: 2015 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS) (2015). https://doi.org/10.1109/iris.2015.7451599
3. Acarman, T., Ozguner, U.: Intelligent cruise control stop and go with and without communication. In: Proceedings of American Control Conference, pp. 4356–4361, Minneapolis, MN, USA, June 2006
4. Jia, D., Lu, K., Wang, J.: A disturbance adaptive design for VANET enabled vehicle platoon. IEEE Trans. Veh. Technol. **63**(2) (2014)
5. Wang, L., Guo, G.: Control over medium-constrained vehicular networks with fading channels and random access protocol: a network system approach. IEEE Trans. Intell. Transp. Syst. **64**, 3347–3357 (2015)
6. Guo, G., Yue, W.: Hierarchical platoon control with heterogeneous information feedback. IET Control Theory Appl. **5**(15), 1766–1781 (2011)
7. Guo, G., Yue, W.: Autonomous platoon control allowing range limited sensors. IEEE Trans. Veh. Technol. **61**(7), 2901–2912 (2012)
8. Naus, G.J.L., Vugts, R.P.A., Ploeg, J., van de Molengraft, M.R.J.G., Steinbuch, M.: String-stable CACC design and experimental validation: a frequency domain approach. IEEE Trans. Veh. Technol. **59**(9), 4268–4279 (2010)
9. Jiau, M.K., Huang, S.C.: Multimedia services in cloud based vehicular networks. IEEE Intell. Transp. Syst. Mag. (2015)
10. Kita, E., Sakamoto, H., Takaue, H., Yamada, M.: Robot vehicle platoon experiment based on multileader vehicle following model. In: Proceedings of Computing and Networking (CANDAR) (2014). https://doi.org/10.1109/candar.2014.30
11. Saxena, A., Li, H., Goswami, D., Belagal, C.: Math design and analysis of control strategies for vehicle platooning. Intell. Transp. Syst. (ITSC) (2016). https://doi.org/10.1109/itsc.2016.7795803
12. Guo, X.G., Liao, F., Teo, R.S.H.: String stability of heterogeneous leader following vehicle platoons based on constant spacing policy. In: Intelligent Vehicles Symposium (IV). IEEE (2016). https://doi.org/10.1109/ivs.2016.7535473
13. Derec, P.U., Chavana, A.A., et al.: Performance analysis of AODV and DSDV routing protocol in MANET and modifications in AODV against black hole attack. In: Elsevier 7th International Conference on Communication, pp. 835–844 (2016)

Unconventional Prediction Algorithm for Quick Route Convergence and Stability in MANET

Mehajabeen Fatima, T. K. Bandopadhyay and Roopam Gupta

Abstract Mobile Ad hoc Network (MANET) is infrastructureless network. Data between source and destination can be communicated through intermediate nodes. Nodes can move randomly and there is no central administration in ad hoc networking. Connectivity with corresponding nodes can be lost due to high mobility, battery power consumption, traffic, and node depletion. This results in repeated route failure. This intrudes the node association and degrades the network's ability to offer the services to its colleague nodes. This makes route maintenance difficult, reduces the stability of the network, and hampers the flow of data. Information related to the route collapse cause can be used to improve the stability of the route. On this basis, Advance AODV (A-AODV) is proposed based on unconventional **prediction algorithms without and with fuzzy logic** to converge the route quickly for continuos data transmission. In this paper, the modifications have been made to the Ad hoc on-demand distance vector (AODV) route maintenance mechanism. Simulations are performed on qualnet 5.0. This paper covers comparative performance analysis of AODV and A-AODV techniques in reference of mobility and traffic. The results illustrated that the route stability can be enhanced through fuzzy prediction algorithm.

Keywords MANET · Cross-layer approach · Route failure · Route convergence Prediction algorithm · Route stability

---

M. Fatima (✉)
SIRT, Bhopal, India
e-mail: mehajabeen.fatima@gmail.com

T. K. Bandopadhyay
BIST, Bhopal, India
e-mail: bando_tushara@gmail.com

R. Gupta
RGPV, Bhopal, India
e-mail: roopamgupta@rgtu.net

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_43

1 Introduction

The mobile computing and communication devices has increased due to developments in software and processing speed of the devices [1]. Laptops, cell phones, smart phones, tablets are considered to generate temporary networks or an ad hoc network just for instantaneous communication without any external involvement [2, 3]. An ad hoc network can be considered as the cooperative engagement of collection of mobile nodes without any control point. Each node acts as a router, mobile nodes can join and leave arbitrarily in the network and can interact dynamically [4]. Ad hoc network mobile nodes can be made compatible with regular Internet. For this, standard network traditional architecture is implemented on all these nodes. Only adjacent layers can communicate with each other in traditional architecture. This is not sufficient to manage with dynamics of ad hoc network. Thus a new architecture is required. Hence, the cross-layer design can be an alternative protocol stack. Cross-layer design breaks hierarchy of traditional network design and this blurs the boundary between two adjacent layers [5]. The information can be exchanged among different layers of the protocol stack in the cross-layer approach [6].

The cross-layer protocol suite is acting as adaptive protocol suite. This can adapt itself to the characteristics of the network [7]. The important part of the adaptive protocol suite is the integration of mobility management schemes. To solve mobility management problems, several different routing protocols have been designed for ad hoc networks. Popular protocols are DSR (Dynamic Source Routing, 1996) and AODV (ad hoc On-Demand Distance Vector Routing, 1999). The AODV performance is better than DSR for mobility, thus AODV is selected for work.

2 AODV and Motivation

AODV has been published as an experimental RFC 3561 in the year 2003 [8]. AODV operation is divided into three parts (a) Route establishment, (b) Route maintenance and (c) Network connectivity. Route can be establishment through Route Request message (RREQ) and Route Reply message (RREP). Source checks its memory for the availability of route whenever it requires [9]. When a node receives a RREP message and sends this message along the reverse path back towards the source node [10]. Then the route will be established and acts as an active route. Hello Message (HM) and Route Error message (RERR) maintains the route. Each active node broadcast HM periodically [11]. When a node does not receives any message from active nodes for some preset period of time then that node is assumed to be no longer available. Then a node generates a RERR message. RERR message propagate to source node through each of its previous. Then the source node restarts the new route discovery procedure. AODV performance is quite good but still has many challenges to resolve. The some challenges of AODV are, Power management [12–16], Security [13, 14, 19], Scalability [15–20], Flooding used for route discovery and local route repair [18–24], Delay in finding of route [18, 19], **Route maintenance** [18], Route search failure [18]. Route maintenance is a major issue of AODV. Thus we pick up route maintenance problem of AODV. Following are the problems arise during route maintenance.

- Periodic broadcast of HM, increases control packets in the network.
- Frequent route collapse because of high mobility and constraint battery power.
- Route collapse makes system unstable.
- Difficult route convergence.
- The route collapse incurred extra cost in terms of RREQ flood, RREP messages, RERR, HM.
- Power consumption will also increases due to frequent route failure.

Due to above issues, modifications are proposed in available AODV route maintenance procedure.

3 Alteration in AODV

Route detection and maintenance are the units of a routing protocol. When sending or forwarding a packet to any destination, route maintenance is used to detect if the network topology has changed. Each node propagating the packet to the next hop on the route is responsible for finding its connectivity to the next hop. To improve performance of the network, AODV is altered to 'Advanced AODV (A-AODV)' for route protection and maintenance in MANET. In A-AODV, route is repaired on highest priority before route breakdown. Broadcast of new maintenance control messages in small area of network reduces overhead. The convergence of the route becomes better through reduction of overhead. The reduction in flooding of control packets reduces battery power consumption as well as repeated route collapse rate. Thus, the route should be maintained locally in advancement of route failure. It can be implemented by considering following points.

- In AODV, only previous node IP address is stored in routing table. The previous and postcursor active node IP address can be saved in routing table through RREQ and RREP messages respectively.
- New control packets can be added. Predict the node serious condition.
- In AODV, Surrounding nodes does not save node IP address after receiving HM. It is suggested that the IP address of active nodes should be saved by inactive surrounding nodes.
- Quick route convergence can be accomplished with route discovery using prediction algorithms without and with fuzzy logic.
- Cross-layer architecture is used for exploitation of information.
- Thus, a novel route discovery method advanced AODV (A-AODV) is proposed for route maintenance of AODV in MANET.

Table 1 Each node entry in routing table after reception of RREQ

Node	Dest	Hop count	Next node
A	–	–	–
B	A	1	A
C	A	2	B
D	A	3	C

Table 2 Each node entry in routing table after reception of RREP

Node	Dest	Hop count	Next node	Forward/Reverse route
A	D	3	B	F
B	A	1	A	R
	D	2	C	F
C	A	2	B	R
	D	1	D	F
A	A	3	C	R

In A-AODV, the route is explored through flooding of RREQ. Consider mobile nodes A, B, C, D, E, F, G, H, I, J are present in a network. Consider A be the source node, Node D be the targeted node and other nodes are intermediate nodes. The routing table does the entry of reverse route after receiving RREQ as shown in Table 1. Table 1 represents reverse route entry in routing table at each node. If a node has route to targeted node or itself targeted node then that node will send RREP to the source node and thus the route is established. The nodes receiving RREP, update their routing table as given in Table 2. The forward route is also saved in each node routing table as shown in Table 2. In Table 2, 'F' denotes forward route entry and 'R' denotes reverse route entry.

The established route is taken as active path and nodes participating in data communication are called active nodes. Let Nodes A, B, C, and D are active nodes and forms active route. Active nodes broadcast its IP address through HM. The inactive neighbor nodes receive HM and do the entry of an IP address of active nodes in its routing table. A link could be broken if one of the active nodes goes out of coverage area or it dies. Hence a limit is decided for these, if these parameters traverses this limit then the node is said to be serious node and the route acts as a serious route. All active nodes call Speed (S), received signal strength (RSS) and residual battery power (RBP) value at the network layer and this is the cross-layer approach.

The active node initiates checking whether the node is in serious situation or not. Data is transmitted as before if node is not serious. If a node is going to be unreachable or when it become serious then the serious node broadcast a Hello warning message (HWM) to warn the nodes that the path is about to break after some time. Assume node B be a serious node, node A acts as previous active node and node C acts as a next active node of serious node B as written in Table 3. The previous node, next node IP of serious node are required to change if a node becomes serious and it is highlighted in Table 4.

Table 3 Node name in reference of serious node

Node name in reference of node B	
Node A	Previous node
Node B	Serious node
Node C	Next node
Node D	Targeted node
Node E, F	Neighbor node (inactive)

Table 4 Routing table entry at each node after advance route repair

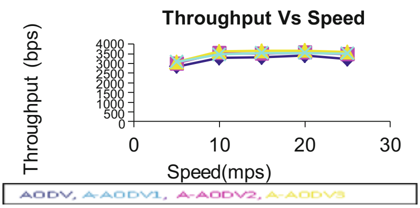
Node	Dest	Hop count	Next node	Forward/Reverse route
A	D	3	***B***	F
B	A	1	A	R
	D	***2***	C	F
C	A	2	***B***	R
	D	1	D	F
D	A	3	C	R

Warning message consists of addresses of previous and post cursor node which are participating in the flow of packets. Neighbor nodes listen the warning message and they will check the broadcast warning node message address in its routing table. If it is having the addresses of both active neighbors then it will reply to serious node by sending warning reply. The node which replies first will participate in new route. Let node E replies first, the serious node sends service replicate message to the previous, next and new node from whom warning reply arrived. The service replicate message consist of logical address of new node (node E), previous and next active node. The service replicate message is broadcast to provide information about address of new node. As soon as service replicate message arrived, node A will replace next node IP entry, node C will replace previous node IP entry in their routing table and node E update its routing table. This sets an alternate hop option in the routing table for previous or next active node and data can be transmitted via this new node without disturbing rest of the path.

Three techniques of A-AODV using above mentioned concept are designed in such a way that the serious route switch to an alternate route in advance before route failure. Techniques are as follows

- A-AODV1: In this method, local network connectivity and route maintenance is accomplished through HM. Route discovery employs preventative route repair is same as discussed in above section.
- A-AODV2: In this method, HM is not used for local connectivity. HM is employed for route maintenance only. This conserves the battery power of inactive nodes which can be exploiting later. Route discovery employs preventative route repair same as discussed in above section.
- A-AODV3: In this approach, the route discovery enabling route repair is same as discussed in A-AODV section. HM is broadcast by active nodes only. Path is

Fig. 1 Throughput versus speed



established through RREQ and RREP. If the node is in serious condition then HWM is generated but this HWM is not generated instantaneously. HWM generates after an interval. This interval is decided on the basis of fuzzy logic. The fuzzy input data activates rules in the fuzzy inference engine to calculate the result. Thus 27 rules are mapped to calculate the warning interval.

4 Performance Analysis and Discussions

The parameters that define a networking context and that should be considered during protocol design, simulation and comparison include:

1. Mobility: Speed and pause time.
2. Traffic: Number of sources, Node density, Data rate, Packet interval, Packet size etc.
3. Scalibility: Node density, Network Size, Number of users, traffic.
4. Node density: Number of nodes, Network size.

This section demonstrates a performance analysis of the standard AODV, A-AODV and techniques running on an identical scenario. Throughput, latency (delay and jitter), residual battery power and route collapse rate during simulation time have been analyzed. The scenario was run for 900 s. for links of 20 and 100 node density. S Qualnet simulator is used for making virtual ad hoc environment. The A-AODV and its techniques are investigated for varying speed and network Size keeping other parameters constant.

There will be increase in traffic with the increase in packet rate, node density, and mobility. The network will become prone to congestion in AODV. This may increase the packet drop rate and thus data packets reception will be decreased. This results in decreased throughput. In A-AODV, an alternate route is searched when route becomes serious and diverts the traffic on another route resulting in less congestion per node. Thus, congestion reduces and therefore throughput and received packets are increased in comparison of AODV as demonstrated from Figs. 1 and 2.

If traffic increases, congestion will also increase. This places the packets in queue and increases the time taken by packet to reach to targeted node. Thus jitter is increased as illustrated in Figs. 3 and 4. The routing protocol with less route collapse rate is most desirable. More packets will be propagated when traffic increases.

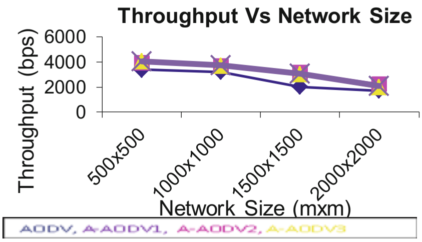


Fig. 2 Throughput versus network size

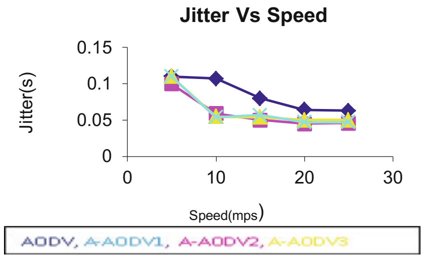


Fig. 3 Jitter versus speed

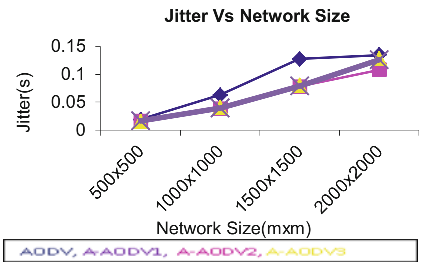


Fig. 4 Jitter versus network size

Route may collapse repeatedly during the transmission of data due of movement of intermediate nodes from coverage area or may be because of node battery failure. Utilization of data related to the route collapse cause in routing could reduce the route collapse probability and improve the network performance. The route collapse cause data can be exchanged through use of cross-layer architecture.

In A-AODV, mobility and limited battery power information are used for prediction of likelihood of route collapse.

The predicted serious situation of route, A-AODV reactively maintains the route and reduces route collapse rate. Therefore, route collapse rate is high in AODV as compared to techniques of A-AODV as shown in Figs. 5 and 6. More traffic results due to frequent route collapse. This in-turn results in the flow of more control packets for route restore in AODV. This increases battery power consumption and as a consequence of this, remaining battery power of a node will be less in AODV as compare to proposed algorithm as shown in Figs. 7 and 8. The HM sent and received control traffic is also reduced in proposed techniques.

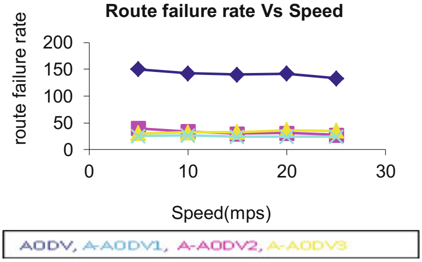


Fig. 5 Route collapse rate versus speed

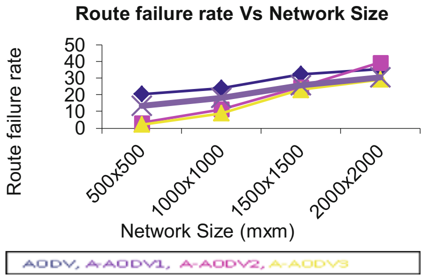


Fig. 6 Route collapse rate versus network size

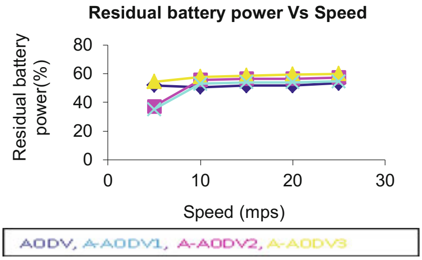


Fig. 7 Residual battery power versus speed

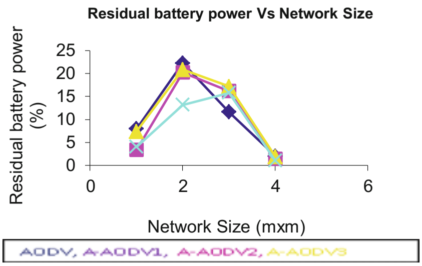


Fig. 8 Residual battery power versus network size

5 Conclusions

The route maintenance mechanism of basic AODV has been revised and named as A-AODV. It is proposed to converge the route quickly through unconventional prediction algorithm without and with fuzzy logic. For this, the warning is broadcast if an active route is predicted in serious condition. The route is serious if it has the possibility of route collapse. The cross-layer architecture is used for prediction of route collapse possibility. The route is switched to another path before route collapse through an innovative algorithm. This improves the stability of the route. The HM flooding is employed for route maintenance only. This not only reduces overhead, but it increases remaining battery power also. This saves battery of inactive nodes. When these nodes appear in active route later, they can utilize the unused battery. This can reduce the possibility of route collapse. The performance of AODV and A-AODV techniques are investigated for mobility and network size in this paper. It is apparent from result analysis that the route collapse rate is reduced which enhances the route stability. The other performance parameters like throughput increased, latency, Overhead (HM sending and receiving rate) are reduced. This improves the performance of the network.

References

1. Sheri, P.K., Jayakkanavar, M., Harshavardan, T.R.: Implementation of AODV protocol for effective disaster management system. s.l. Int. J. Adv. Technol. Eng. Res. (IJATER) **2**, 50–58 (2012)
2. Mohapatra, P., Li, J., Gui, C.: QoS mobile ad hoc network. IEEE Wirel. Commun. **10**, 01–33 (2003) (Sep 2000)
3. Gerla, M., Lee, S.-J.: AODV-BR: backup routing in ad hoc networks. s.l. IEEE WCNC **3**, 1311–1316
4. Kumar, K., Gulati, M.K.: QoS routing protocols for mobile ad hoc networks: a survey. Int. J. Wirel. Mob. Comput. **5**, 107–118 (2012)
5. Motani, M., Srivastava, V.: Cross-layer design: a survey and the road ahead. s.l. IEEE Commun. Mag. **43**, 112–119 (2005)
6. Preetha, K.G., Unnikrishnan, A., Jacob, K.P.: A probabilistic approach to reduce the route establishment overhead in AODV algorithm for MANET. s.l. Int. J. Distrib. Parallel Syst. **3**, 207–214 (2012)
7. Prochart, G., Weiss, R., Schmid, R., Kaefer, G.: Fuzzy-based support for service composition in mobile ad hoc networks. IEEE Int. Conf. Pervasive Serv. 379–384 (2007)
8. Perkins, C.E., Royer, E.M.: An implementation study of the AODV routing protocol. In: Proceedings of the IEEE Wireless Communications and Networking Conference, Los Angeles, Sept 2000
9. Mhala, N., Choudhari, N.K.: An implementation possibilities for AODV routing protocol in real world. Int. J. Distrib. Parallel Syst. (IJDPS) **01**, 118–127 (2010)
10. Chakeres, I.D., Belding-Royer, E.M.: The utility of HMs for determining link connectivity. In: International Symposium on Wireless Personal Multimedia Communications (WPMC), Honolulu, Hawai, vol. 2, pp. 504–508, Oct 2002
11. Preetha, K.G., Unnikrishnan, A., Jacob, K.P.: A probabilistic approach to reduce the route establishment overhead in AODV algorithm for MANET. Int. J. Distrib. Parallel Syst. **3**(2), 207–214 (2012)
12. Larsen, K.G., Pettersson, P., Yi, W.: Uppaalin in a Nutshell. Int. J. Softw. Tools Technol. Transf. **1**(1–2), 134–152 (1997)
13. Choi, B.J., Shen, X.: Adaptive exponential beacon period protocol for power saving in delay tolerant networks. In: IEEE ICC Proceedings, pp. 1–6. IEEE Communications Society (2009)

14. Guoqing, L., Zhaowei, Q.: Unidirectional link problem in AODV routing protocol. In: 2nd IEEE International Conference on Broadband Network & Multimedia Technology (IC-BNMT'09), pp. 18–20 (2009)
15. Ji, Z., Yu, W., Liu, K.J.R.: A belief evaluation framework in autonomous MANETS under noisy and imperfect observation: vulnerability analysis and cooperation enforcement. IEEE Trans. Mob. Comput. **9**(9), 1242–1254 (2010)
16. Liu, G., Wong, K.J., Lee, B.S., Seet, B.C.: PATCH: a novel local recovery mechanism for mobile ad hoc networks. In: 58th IEEE Vehicular Technology Conference, Orlando, USA, vol. 5, pp. 2995–2999, 6–9 Oct 2003
17. Banerjee, A., Dutta, P.: Fuzzy controlled localized route-repair [FLRR] for on demand routing protocols in mobile ad hoc networks. Int. J. Comput. Appl. **7**, 5–10 (2010)
18. Jain, J., Gupta, R., Bandhopadhyay, T.K.: Scalability enhancement of AODV using local link repairing. s.l. Int. J. Electron. 1–15 (2013)
19. Lee, S.-J., Belding Royer, E.M., Perkins, C.E.: Scalability study of the ad hoc on demand distance vector routing protocol. Int. J. Netw. Manag. **12**, 97–114 (2003)
20. Goff, T., Abu-Ghazaleh, N.B., Phatak, D.S.: Preemptive routing in ad hoc networks. In: Proceedings of ACM SIGMOBILE, Seventh Annual International Conference on Mobile computing and networking, Rome, Italy, pp. 43–52, 16–21 July 2001
21. Ren, Z., Su, J., Guo, W.: A cross-layer AODV routing protocol. In: IEEE International Conference on Mechatronics & Automation, Niagara Falls, Ontario, Canada. pp. 2150–2155, 29 July–1 Aug 2005
22. Vuran, M.C., Akyildiz, I.F.: XLP: a cross-layer protocol for efficient communication in wireless sensor networks. IEEE Trans. Mob. Comput. **9**, 01–14 (2010)
23. Manickam, J.M.L., Shanmugavel, S.: Fuzzy based trusted ad hoc on-demand distance vector routing protocol for MANET. In: Third IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (Wimob), New York USA, pp. 110–119 (2007)
24. Chen, X., Faloutsos, M., Krishnamurthy, S.V.: Power adaptive broadcasting with local information in ad hoc networks. In: 11th IEEE International Conference on Network Protocols, Atlanta, USA, pp. 243–261, 4–7 Nov 2003

Analysis on Logical Key Hierarchy and Variants for Secure Group Communication

Aparna S. Pande, Yashwant Joshi and Manisha Y. Joshi

Abstract In secure group communication for applications such as pay-per-view, teleconferencing, and video conferencing, key management method with logarithmic computation is expected along with storage and bandwidth efficiency. Major aspects with key management are formation of group key with initial group members and updating the group key with any change in number of group members. Rekeying is mandatory due to group join/leave operations or periodic rekeying. Logical Key Hierarchy has logarithmic rekeying cost. This paper explores Logical Key Hierarchy (LKH), LKH variants with centralized and distributed approach are analyzed and presented. This paper explores their way of achieving and persisting logarithmic cost for secure group communication, to decide the suitable method for a current application.

Keywords Group key management · Secure multicast · Rekeying · LKH · OFT OFC · IHC · SD-LKH · Distributed LKH · Diffie–Hellman LKH · Flat key table LKH · Mykil

1 Introduction

In group communication, multiple receivers receive same copy of transmission at the same time when sender has sent just one copy of data. In secure group commu-

A. S. Pande (✉)
Department of Computer Science & Engineering, SGGSIE&T, Nanded, Maharashtra, India
e-mail: aparnajoshi10@yahoo.com

Y. Joshi
Department of Electronics & Telecommunication Engineering, SGGSIE&T, Nanded, Maharashtra, India
e-mail: yvjoshi@sggs.ac.in

M. Y. Joshi
Department of Computer Science & Engineering, MGMCE, Nanded, Maharashtra, India
e-mail: joshi_my@mgmcen.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_44

nication, only intended and authorized users (receivers and sender/s) should send and receive data and others should be prohibited to take part in communication, so group key is used for all communication. With change in group dynamics due to join or leave operation, group key has to be changed or regenerated to ensure backward and forward secrecy. If “T” denotes current time, then backward secrecy is if a user is joining at time “T”, then user should not get access to all communication before time “T”, but henceforth should participate in group communication. Forward secrecy means, if a user is leaving at time “T”, then it should not participate in further communication after time “T”. Change in group key is called as rekeying. Rekeying happens due to the result of join/leave operation or periodic time interval (batch rekeying). In periodic rekeying, group key is changed only, if the set period has reached to its limit, before that any number of join or leave will not alter the group key. It is also known as batch rekeying as within set period all join/leave are processed as a batch operation.

Lets assume that there is a trusted server to store membership information (for group access). When a user sends join request to the group, group server using authentication protocol mutually authenticates user. With the list of all authenticated and accepted members, group server distributes group key to group. To ensure forward/backward secrecy, server has to form new group key and convey it to all members whenever there is a change in group dynamics, so always only current group members are participating in group communication.

From [1–5], key distribution scheme performance evaluation parameters are given in Table 1. In hierarchical tree approach, a rooted tree, where each leaf node denotes a user is constructed in bottom-up manner. The server authenticates user and keeps each user key. Server is responsible for generation and distribution of group key. Hierarchical tree approach achieves [1, 5] logarithmic storage cost for every user and logarithmic rekeying transmissions are required for rekeying. The motivation behind this paper is to explore LKH and to provide a way for each application, how to persist logarithmic cost (Storage, communication or rekeying cost). Among various key management methods [1, 5–7], in this paper, we defined working of each LKH variant method only, how it achieves logarithmic cost, for which application, and based on its core functioning method one can select among LKH variants a suitable method for the current purpose.

2 Logical Key Hierarchy

Wallner et al. and Wong et al. [1, 8] proposed a logical key hierarchy (LKH). In [4], a survey on LKH is presented with performance evaluation parameters for LKH and other counter key management methods. In LKH method, group controller (GC) maintains a Key tree. In Fig. 1, key tree has U-nodes (which indicates users as leaf and has its individual key) and K-nodes (either subgroup key or group key). All group members share and participate in group communication with group key. Each user keeps user key, group key, and all intermediate key from current node to along

Table 1 Key distribution scheme performance evaluation parameters

Parameter		Description
Efficiency	Initial keying	Formation of group with initial access list
	Rekeying	Reformation of group after join/leave semantics
Scalability		Whether there is either no change in performance or acceptable change in performance with increase in number of users and operations
Computation requirement		Computational resources required for server and users
Storage requirement		Storage requirements at server side and user side
Stateness		Stateful if new keys are encrypted with old keys for distribution

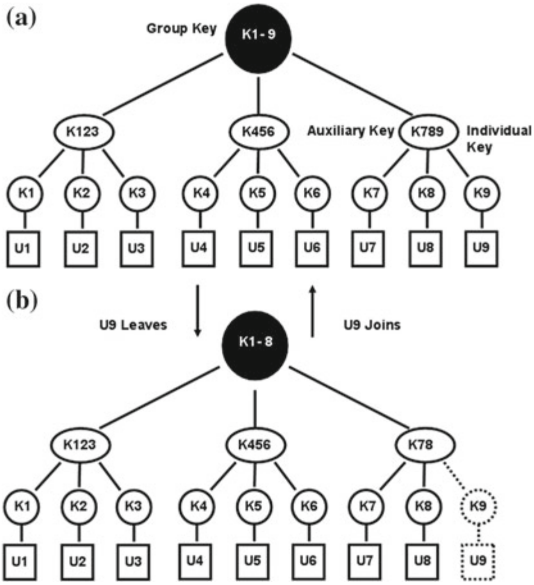


Fig. 1 Key tree of degree 3

the path to the root (key path/external path). In Fig. 1, individual keys are K1–K9, subgroup keys are K123, K456, and K789, and group key is K1–9. When there is any change in group due to join/leave group semantics, all keys that is known to user (key path) needs to change.

Procedure Join: If user U9 joins the group, new group keys k1–9 is generated to replace previous group keys k1–8 and new subgroup key k789 will replace previous subgroup key k78. LKH is stateful in nature, and for secure distribution of new keys to authenticated group members, it uses previous keys to encrypt new keys. When user U9 joins, old group keys k1–8 is used to encrypt new group k1–9 and new group key is distributed to all group members. Key k78 is used to encrypt k789 and distributed to users along the path only. Keys k1–9, k789 are encrypted with k9 and is

Table 2 LKH performance evaluation parameters

Parameters		Description
Backward and forward secrecy		Maintains backward and forward secrecy
Collusion free		Secure against collusion of evicted users as old keys are not used to encrypt new keys
Storage cost	Server side	For server it is 2sN, as it stores all keys in the key tree
	User side	For each user it is O(log N), as user stores the keys on its key path
Join/leave operation cost	Server side	O(log N)
	User side	O(1)
Stateness		Stateful
Approach		Centralized, key based and bottom-up
Rekeying		After each join/leave operation (no batch or periodic)
Key tree		Balanced tree is required to achieve logarithmic performance

given to user U9 only. The server multicasts the following messages to communicate new keys to the intended users.

$$\begin{aligned} &Key\ Server \rightarrow All{:}Enc\{k1{-}9, k1{-}8\},\\ &Key\ server \rightarrow U7, U8{:}Enc\{k789, k78\},\\ &Key\ server \rightarrow U9{:}Enc\{k1{-}9|k789, k9\}. \end{aligned} \tag{1}$$

Procedure Leave: If user u9 leaves the group, change keys k1–9 with k1–8 and k789 with k78. In leave operation, old keys cannot be used to encrypt new keys as sold keys are known to the revoked user. To ensure forward secrecy, encrypt keys with its child keys. As keys are not known to revoked users, it avoids collusion attack.

$$\begin{aligned} &Key\ Server \rightarrow U1, U2, U3{:}Enc\{k1{-}8, k123\},\\ &Key\ Server \rightarrow U4, U5, U6{:}Enc\{k1{-}8, k456\},\\ &Key\ Server \rightarrow U7, U8{:}Enc\{k1{-}8, k78\}\\ &Key\ Server \rightarrow U7{:}Enc\{k78, k7\}\\ &Key\ Server \rightarrow U8{:}Enc\{k78, k8\}. \end{aligned} \tag{2}$$

In Table 2, parameters for performance evaluation of LKH are given. For a group of size N with key tree of "d" degree, LKH has O logdN communication cost (Table 2).

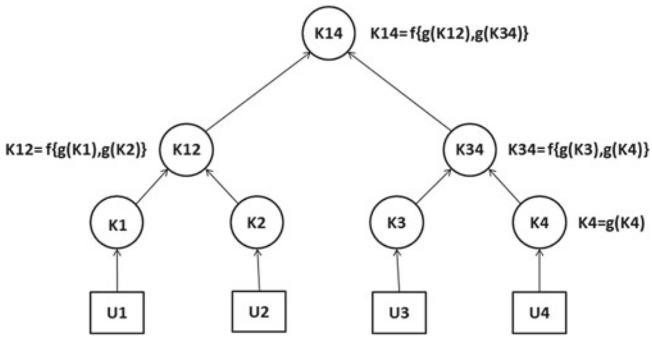


Fig. 2 One-way function chain tree

3 LKH Variants

In this section, LKH variants are presented. Among all key management methods [1, 4, 5, 9, 10], in this paper, hierarchical methods working with centralized or distributed approach and which are close to performance of LKH are presented. Each hierarchical key distribution method works in a different manner. That difference in maintaining hierarchical key tree is studied and presented. LKH and its variants are analyzed to know the working of each method, its core function, computation, and communication way along with its differences from each other to know which suitable method is for an application. Table 3 summarizes comparison of these variants with LKH.

i. **One-Way Function Tree (OFT)**

Balenson et al. [11] proposed this variant One-Way Function Tree. In OFT, like LKH server maintains a binary tree where interior node keys are derived from its child keys. To derive the node keys, "g" a special one-way function, "f" a mixing function and exclusive or is used. In Fig. 2, each interior node key Kv is derived using its left and right child keys. To compute node key Kv, apply one-way function "g" on each child and mix it using mixing function "f" and bitwise exclusive XOR.

With OFT though f(Kv) is known and it is hard to recover Kv because of one-wayness of "f" function thus f(Kv) is also known as blinded key of Kv. Every user has individual node key and set of blinded keys (keys of siblings of nodes on key path). Group key is computed in bottom-up manner using node keys and set of blinded keys. In Fig. 2, user U4 has individual key K4, blinded keys f(K3), and f(K12). With f(K3) and K4, it can derive K34 and with f(K12) and K34 and it can compute group key. Thus, update node keys and set of blinded keys on key path after each join/leave operation.

In OFT, new member is added, by splitting existing node near to the root and attaching new one to right and existing one to left. Update node keys and corresponding set of blinded keys to ensure backward secrecy. OFT is bandwidth

Table 3 Performance evaluation of LKH and variants

Method	Approach	Server comm. cost	Join rekeying cost	Leave rekeying cost	No. of rounds	Suitable for application	Core of method
LKH	Centralized, bottom-up	O(logN)	d + 1	2d	–	With large revoked users	Use of key tree
OFT	Centralized, bottom-up	O(logN)/2	d + 1	d + 1	–	Low bandwidth, large revoked users	Use of one-way function and blinded keys
OFC	Centralized, bottom-up	O(logN)	logdN +1	(d − 1) logdN	–	More leave user than add, low communication cost	Use of pseudo random generator
IHC	Centralized, bottom-up	O(logN)	1 + logdN	(d − 1) logdN	–	More leave user than add, low communication cost	Use of one-way hash function and works with higher degree trees
SD-LKH	Centralized, bottom-up	O(logN)	logdN +1	(d − 1) logdN	–	More leave user than add, low communication cost	Use of simple XOR applying distribution of difference
Flat key table LKH	Centralized, bottom-up	(2I + 1)K	2IK	2IK	–	Low communication cost, scalable	Maintains a flat table with reduced no of keys held by KDC
Distributed LKH	Distributed, bottom-up	log2Nd	–	–	log2N	Distributed and minimal keys held by user	Distributed mutual key agreement, inherently parallel

(continued)

Table 3 (continued)

Method	Approach	Server comm. cost	Join rekeying cost	Leave rekeying cost	No. of rounds	Suitable for application	Core of method
Diffie–Hellman LKH	Distributed, bottom-up	log2N + 1	–	–	log2N	Distributed and minimal keys held by user	Distributed Diffie-Hellman, contributory group key computation
Hybrid	Hybrid, key based, bottom-up	O(logN)	Batch rekeying	Batch rekeying	–	Large groups with frequent join/leave	Distributed at top level and centralized at local

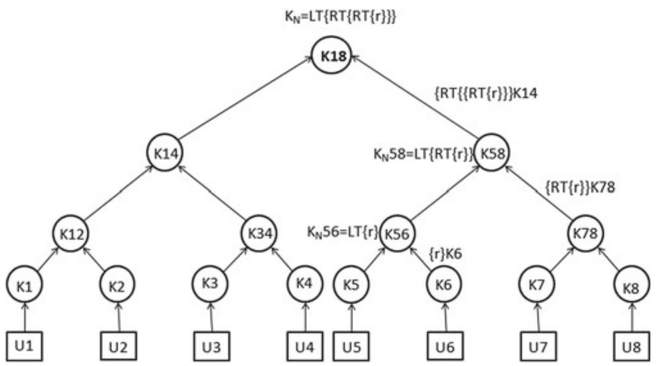


Fig. 3 One-way function tree

efficient when compared to LKH as it does not need to distribute all keys on the path.

One can derive group key and other blinded keys having few blinded keys and node keys. In OFT, when a member is evicted, then attach sibling of evicted member at its location and assign new key to it, if it is a leaf node and if evicted member is a subtree node, then shrink tree move close to the root, and assign new key to updated members to make it secure against collusion.

ii. **One-way Function Chain Tree (OFC)**

Canetti et al. [12] proposed One-Way Function chain Tree. In OFC, like OFT node keys are derived from its child keys. To derive node keys as shown in Fig. 3, [13], a pseudo-random generator is used, functions LT(x) and RT(x) are output of PSRG(x). To make it secure, each node key is encrypted using its siblings KEK.

$$PSRG(x) = LT(x)\,RT(x), and\; |LT(x)| = |RT(x)| = |x| \tag{3}$$

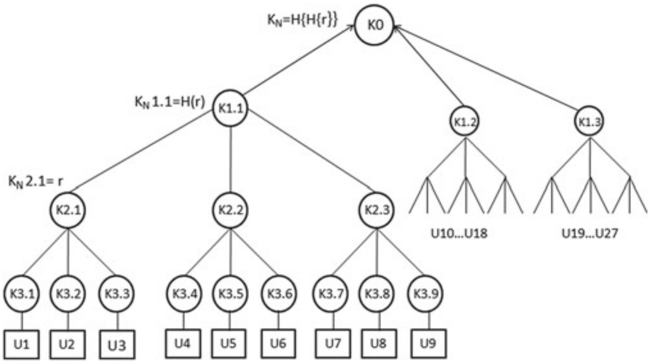


Fig. 4 Iterated hash chain tree

OFC is efficient in communication overhead than LKH in leave user operation only. When a user leaves the group, new value r is assigned to the sibling of the evicted user, rp(u) = r and for rest of the nodes assign rp(v) = RT(r), p(v) is parent of node v. New node keys are pseudo-random function of LT(rv) in bottom-up manner.

In Fig. 3, when user u5 left, assign values, r encrypted with K6 to node n56, RT(r) encrypted with K78 to node n58 and RT (RT(r)) encrypted with K14 to node n0. New KEKs are derived as

$$\begin{aligned} KN56 &= LT(r), \\ KN58 &= LT\,(RT(r))\, and \\ KN &= LT\,(RT\,(RT(r))). \end{aligned} \tag{4}$$

It ensures forward, backward secrecy, and it is secure against collusion of evicted users.

iii. **Iterated Hash Chain Tree (IHC)**

Zhu [14] proposed IHC to achieve performance of OFC with higher degree tress. OFC supports binary tree only. IHC uses a one-way hash function H [13] which derives node keys in bottom-up manner applying iteration of hash function from leaf to root. In Fig. 4, from a ternary tree, when user u1 left group, then at random select, a key "r" and with iterated hash function H compute node keys K2.1, K1.1, and K0. Key server sends the following messages to convey computed node keys.

$$\begin{aligned} &Keyserver \rightarrow U2{:}Enc\{\{r\}K3.2\}, \\ &Keyserver \rightarrow U3{:}Enc\{\{r\}K3.3\}, \\ &Keyserver \rightarrow U4, U5, U6{:}Enc\{\{H\{r\}\}K2.2\}, \\ &Keyserver \rightarrow U7, U8, U9{:}Enc\{\{H\{r\}\}K2.3\}, \\ &Keyserver \rightarrow U10{-}U18{:}Enc\{\{H\{H\{r\}\}\}K1.2\}, \\ &Keyserver \rightarrow U19{-}U27{:}Enc\{\{H\{H\{r\}\}\}K1.3\} \end{aligned} \tag{5}$$

If user U1 joins the group, then key server sends messages

$$
\begin{aligned}
&Keyserver \rightarrow U1{:}Enc\{\{r\}K3.1\},\\
&Keyserver \rightarrow U2, U3{:}Enc\{\{r\}K2.1\},\\
&Keyserver \rightarrow U4{-}U9{:}Enc\{\{H\{r\}\}K1.1\},\\
&Keyserver \rightarrow U10{-}U27{:}Enc\{\{H\{H\{r\}\}\}K0\}
\end{aligned} \tag{6}
$$

To compute node keys, KN = Hn(r), where n = logd N − 1. One-way function H provides backward, forward secrecy as well as it is secure from collusion.

iv. **Synchro-difference LKH (SD-LKH)**

Zhu [14] proposed SD-LKH to achieve performance of IHC with simplified XOR, instead of complex one-way function. In SD-LKH, new node keys are computed with distribution of difference from previous keys. The key server at random distributes to all users a differential value D encrypted with users key. In Fig. 4, when user U1 left group, key server selects value D, encrypts with users key, and sends messages

$$
\begin{aligned}
&Keyserver \rightarrow U2{:}Enc\{\{D\}K3.2\}\ Keyserver \rightarrow U3{:}Enc\{\{D\}K3.3\},\\
&Keyserver \rightarrow U4, U5, U6{:}Enc\{\{D\}K2.2\},\\
&Keyserver \rightarrow U7, U8, U9{:}Enc\{\{D\}K2.3\},\\
&Keyserver \rightarrow U10{-}U18{:}Enc\{\{D\}K1.2\},\\
&Keyserver \rightarrow U19{-}U27{:}Enc\{\{D\}K1.3\}
\end{aligned} \tag{7}
$$

Knowing value D, each user computes the new keys with XOR operation [13]

$$
\begin{aligned}
&\mathrm{KN2.1} = \{\mathrm{K2.1}\|\mathrm{D}\},\\
&\mathrm{KN1.1} = \{\mathrm{K1.1}\|\mathrm{D}\},\\
&\mathrm{KN} = \{\mathrm{K0}\|\mathrm{D}\}
\end{aligned} \tag{8}
$$

v. **Flat Key Table LKH**

Waldvogel et al. [13] proposed a flat key table approach for secure group communication. The flat table has 1 TEK and 2w KEK, where w denotes total bit values in member ID. For each possible bit, value 0 or 1, 2 KEKs are assigned to each member ID. Each member holds w + 1 key. In any centralized key management approach, number of keys stored by each member and KMS depends on how the subgroups are formed. Table 4 shows a CFT with group size 16. Each member has 4-bit id. In CFT approach, KMS stores fewer keys than LKH but in LKH and CFT, each member stores the same number of keys. With change in group dynamics due to join or leave, KMS changes keys possessed by that member. Required rekeying messages to update the change in group key are in linear proportion with the bits in group member ID. If group size is 4, then $\log_2$ n + 1 means (log2 4 + 1) = 3 rekeying messages will require updating change.

Table 4 CFT with group size 16

	Traffic encryption key	
ID bit#0	KEK0,0	KEK0,1
ID bit#1	KEK1,0	KEK1,1
ID bit#2	KEK2,0	KEK2,1
ID bit#3	KEK3,0	KEK3,1
	Bit value = 0	Bit value = 1

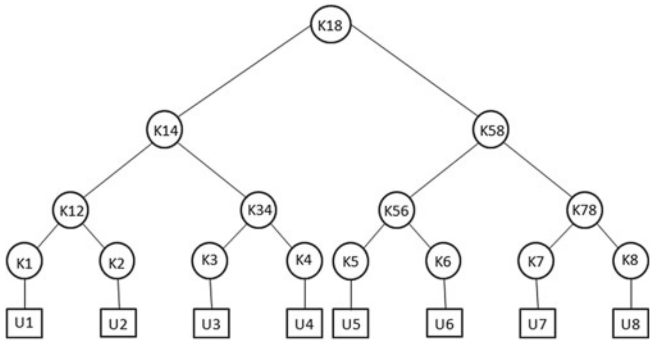


Fig. 5 Distributed LKH tree

This scheme is storage efficient as server stores $2\log_2 n$ keys which is less than LKH. The scheme is not collusion-free, as together evicted members may get total set of keys. In [15], the authors improved performance of CFT, where rekeying messages are independent of group sizes.

vi. **Distributed LKH Tree (D-LKH)**

Rodeh et al. [16] proposed Distributed LKH. Being distributed in nature, there is no centralized group controller like in LKH, thus at a time, all keys are not known to anyone. Hierarchy of keys is generated among users with the help of mutual key agreement. As there is no leader or centralized entity, to set up, all members compute intermediate values independently and in parallel manner. At the final round, all members will compute the same group key. In distributed LKH tree, in Fig. 5, each subtree key is a result of mutual key agreement in users, for subtree key K12, user 1 and user 2 agrees mutually on key K12, user3 and user4 agrees on K34, user 5 and user 6 agrees on K56, and user 7 and user8 agrees on K78. This mutual key agreement continues till all users agree on group key K18. For N users, algorithm takes log2 N rounds to complete.

vii. **Diffie–Hellman LKH**

Kim and Perrig [17] proposed distributed Diffie–Hellman LKH. To derive node keys and Group Key in contributory manner, Diffie–Hellman algorithm is used. In Fig. 5, [13] with Diffie–Hellman algorithm node keys are, Key K12 is result of αK1K2 mod P, αK3K4 mod P computes K34, so is for K56 (αK5K6 mod P), K78

(αK7K8 mod P). This algorithm will continue till group key is computed, K18 (αK14K58mod P.) for group of N users, to complete this algorithm, log2sN rounds are required. Fixed number of rounds indicates that the operations can be done in parallel which minimizes time required to compute group key. It is also scalable as number of group members and required interactions are independent entities.

viii. **Hybrid Tree: Mykil (Multi-Hierarchy-Based Key Distribution)**

Huang et al. [18] proposed Mykil. A combination of Iolus [19], LKH [1, 8], and Kronos [20]. Mykil divides group into subgroups (area) and there is an associated area controller with each area, within each subgroup, hierarchy of keys is formed. At higher level, Mykil is distributed (Iolus) and within each area, it is centralized (LKH), so scope of join/leave operation is within a subgroup. Mykil adopts Batch rekeying thus after each join/leave operation all keys are not updated immediately instead rekeying is postponed till set period interval has not occurred. Mykil tree is balanced as it prunes node after leave user operation. In Mykil group, rekeying rate is constant and is not affected by change in group dynamics and independent of group size.

Notations used in paper and Tables 2, 3, 4

- N is number of group members
- d is depth of tree
- Enc is Encryption
- D is distribution difference
- XOR operation
- α is Diffie Hellman Prime
- PSRG is pseudo random generator
- H is one-way hash function
- g is a special one-way function
- f is a mixing function
- K_i is key of ith member
- I is number of bits in member id
- K is size of key in bits.

4 Conclusion

Logical key hierarchy and its variants with centralized or distributed approach are analyzed; working of each method along with its properties is presented. Comparison of all these variants is given in Table 3. LKH and all variants provide forward, backward secrecy, and all are secure from collusion of evicted users. OFT, OFC, IHC, and SD-LKH are bandwidth efficient. CFT is storage efficient. In distributed environment, distributed LKH and Diffie–Hellman LKH achieves logarithmic performance. Mykil is a hybrid approach and most scalable. Table 3 summarizes communication cost, rekeying cost, and core function of each LKH variant, which will help to find most suitable method for a particular application. Before selecting any of the security solution for an application, one must be aware of total requirements of that application as the best solution for an application may not be for other application.

References

1. Wallner, D., Harder, E., Agee, R.: Key management for multicast: issues and architectures. IETF, RFC 2627 (1999)
2. Yu, X.: Survey on the hierarchical key distribution algorithms for secure multicasting. http://www.comp.nus.edu.sg/~cs5248/surveys/xusecure-key.pdf
3. Kruus, P.S.: A survey of multicast security issues and architectures. In: 21st National Information Systems Security Conference, Arlington, VA, Oct 1998
4. Pande, A.S., Thool, R.C.: Survey on logical key hierarchy for secure group communication. In: IEEE International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Sept 2016, pp. 1131–1136
5. Harney, H., Harder, E.: Logical key hierarchy protocol. draft-harney-sparta-lkhp-sec-00.txt. IETF Internet Draft (1999)
6. Diot, C., Levine, B.N., Lyles, B., Kassel, H., Balensiefen, D.: Deployment issues for the IP multicast service and architecture. IEEE Netw. **14**, 78–88 (2000)
7. Rafaeli, S., Hutchison, D.: A survey of key management for secure group communication. ACM Comput. Surv. **35**(3), 309–329 (2003)
8. Wong, C.K., Gauda, M., Lam, S.S.: Secure group communications using key graphs. IEEE/ACM Trans. Netw. **8**(1), 16–30 (2000)
9. Zhu, S., Jajodia, S.: Scalable group key management for secure multicast: a taxonomy and new directions. http://www.cse.psu.edu/~szhu/papers/survey.pdf
10. Goshi, J., Ladner, R.E.: Algorithms for dynamic multicast key distribution trees. In: Proceedings of ACM Symposium on Principles of Distributed Computing (PODC) (2003)
11. Balenson, D., McGrew, D., Sherman, A.: Key Management for Large Dynamic Groups: One-Way Function Trees and Amortized Initializations. IETF Draft, Aug 2000
12. Canetti, R., Garay, J., Itkis, G., Micciancio, D., Naor, M., Pinkas, B.: Multicast security: a taxonomy and some efficient constructions. In: IEEE INFOCOM, New York, vol. 2, pp. 708–716, Mar 1999
13. Waldvogel, M., Caronni, G., Sun, D., Weiler, N., Plattner, B.: The VersaKey framework: versatile group key management. IEEE J. Sel. Areas Commun. **17/8** 1614–1631 (1999)
14. Zhu, W.T.: Optimizing the tree structure in secure multicast key management. IEEE Commun. Lett. **9**(5) (2005)
15. Joshi, M., Bichkar, R.: A centralized key table based communication efficient group key management protocol. Int. J. Comput. Netw. Inf. Secur. (8), 49–55 (2015)
16. Rodeh, O., Birman, K., Doley, D.: Optimized group rekey for group communication systems. Technical Report 2, Department of Computer Science, Hebrew University (1999)
17. Kim, Y., Perrig, A., Tsudik, G.: Simple and fault tolerant key agreement for dynamic collaborative groups. In: Proceedings of 7th ACM Conference in Computer and Communication Security, pp. 235–241, Nov 2000
18. Huang, J.H., Mishra, S.: Mykil: a highly scalable key distribution protocol for large group multicast. In: IEEE Global Communications Conference (2003)
19. Mitrra, S.: Iolus: a framework for scalable secure multicasting. In: ACM SIGCOMM, pp. 277–288 (1997)
20. Setiya, S., Koussih, S., Jajodiya, S.: Kronos: a scalable group rekeying approach for secure multicast. In: Proceedings of IEEE Symposium on Security and privacy (2000)

Performance Analysis of SLM Technique for PAPR Reduction in OFDM Using QPSK Modulation

Amol B. Kotade, Anil Nandgaonkar and S. L. Nalbalwar

Abstract Multi-carrier communication is a backbone of fourth-generation (4G) mobile communication due to high-data-rate capability. This multi-carrier communication is facilitated using orthogonal frequency-division multiplexing (OFDM). But high peak-to-average power ratio (PAPR) is a serious problem present in OFDM. High PAPR causes huge battery power consumption. To reduce PAPR, selected mapping (SLM) is one of the good technique from the set of scrambling methods. The performance of SLM technique for QPSK-modulated data with phase offsets of $\pi/4$, $\pi/2$, and $3\pi/2$ using $N = 64$, 128, 512, 1024, and 2048 number of subcarriers in MATLAB environment is presented in this article. PAPR results of original OFDM and OFDM with SLM technique are plotted and also presented in tabular form for comparison. We have also investigated the computational complexity of SLM scrambler and descrambler block using Freescale StarCore SC140 architecture simulator.

Keywords Selected mapping · Peak-to-average power ratio · OFDM

1 Introduction

To provide high-data rate with high-spectral efficiency for the end user is a key for evolution of wireless technology. Hence, fourth- and next-generation wireless devices are demanding supercomputing performance with very less battery power consumption. In OFDM, data is transmitted using a number of subcarriers. These

A. B. Kotade (✉) · A. Nandgaonkar · S. L. Nalbalwar
Electronics and Telecommunication Engineering Department, Dr. Babasaheb Ambedkar Technological University, Lonere, Raigad 402103, Maharashtra, India
e-mail: kotade.amol@gmail.com

A. Nandgaonkar
e-mail: abnandgaonkar@yahoo.com

S. L. Nalbalwar
e-mail: slnalbalwar@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_45

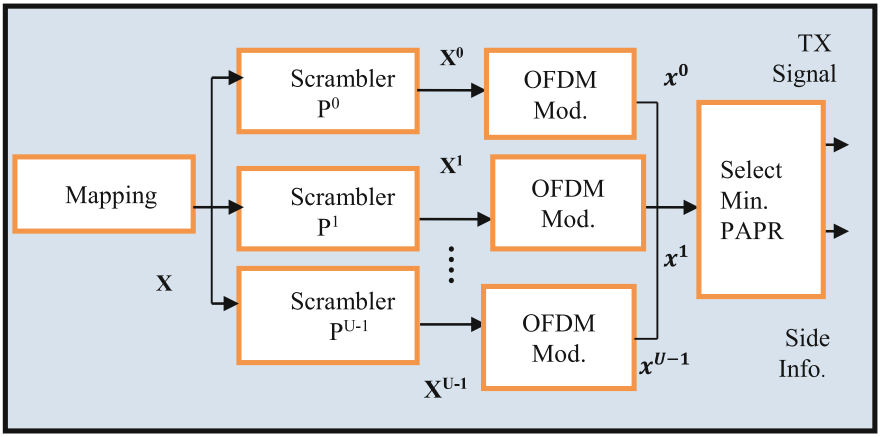


Fig. 1 SLM technique using scrambler

subcarriers are separated orthogonally from each other. When "N" number of various signals of the same phases are combined it causes,

$$\text{Peak Power} = \text{N} \times \text{Average Power of signal.}$$

This results in high PAPR. High PAPR makes communication system power inefficient thereby decreases the battery life of mobile devices. SLM [1] is a signal scrambling technique of PAPR reduction. SLM is observed to be more effective in terms of PAPR reduction than other techniques such as clipping [2], block coding [3], and partial transmit sequence (PTS) [4]. This paper presents the performance analysis of SLM technique. In the following sections, we have explained the implementation of SLM, its performance in terms of PAPR reduction for various number of subcarriers and computational complexity of scrambler/descrambler used in SLM.

2 SLM with Scrambler

2.1 Implementation

The implementation of SLM is shown in Fig. 1. The quadrature phase-shift keying (QPSK) modulated data ($\text{X} = [\text{X}_0; \text{X}_1; \ldots; \text{X}_{\text{N}-1}]$) is produced by mapper block. Here, "N" is the number of subcarriers to be used for data transmission. Output of mapper is given to "U" number of scramblers. The scrambling equation is [5]

$$X^u = X \cdot P^u, \tag{1}$$

where P^u is the scrambler matrix and

$$p_n^u = (+1, -1, +j, -j). \tag{2}$$

For example, let N = 4, X = [0.707, j, −0.707, −j] and the scrambling matrix is p_n^u, then scrambled sequence is given by the following equation:

$$X^u = [0.707, -j, -0.707j, -1]. \tag{3}$$

Thus, SLM scrambler performs a kind of phase rotation of sequence. This scrambled sequence is OFDM modulated to result in x^u. Finally, symbol x^u with lowest value of PAPR is selected for transmission. The scrambler randomizes the data sequences, which reduces the PAPR [3, 4]. More randomness causes a drastic reduction in PAPR along with synchronization. So, scrambling has twin benefit in contemporary communication system design. In SLM there are a total of U number of scrambler blocks, for the reduction of PAPR data that needs to be randomized more. This is achieved by using more scrambling operation. This causes more computational complexities in terms of machine cycle consumption when ported on real-time processor. This indicates that larger PAPR reduction is possible at the cost of increased computational complexity of SLM scrambler. So, we also need to look for machine cycle consumption for given SLM scrambler to meet real-time processing constraint.

2.2 Generation of Scrambling Sequence

The block diagram for generating the scrambling mask is shown in Fig. 2 [6]. The connection polynomial for performing the XOR operation is 1 + D + D15 [6]. Initial values to be stored in delay elements of scrambler is given by 1 + D + D3 + D6 + D8 + D10 + D11 + D14 polynomial [6]. For each incoming data, all delay elements are clocked in order to produce scrambling sequence. The polynomial selection is done based on the performance requirement at the physical layer of a particular wireless technology. These polynomials are different for processing the paging and control channels at that physical layer [6]. In the further sections, we have discussed the PAPR reduction performance of SLM technique for QPSK-modulated data with different phase offsets and a number of subcarriers. Also, computational complexity of SLM scrambler/descrambler in terms of machine cycle is determined.

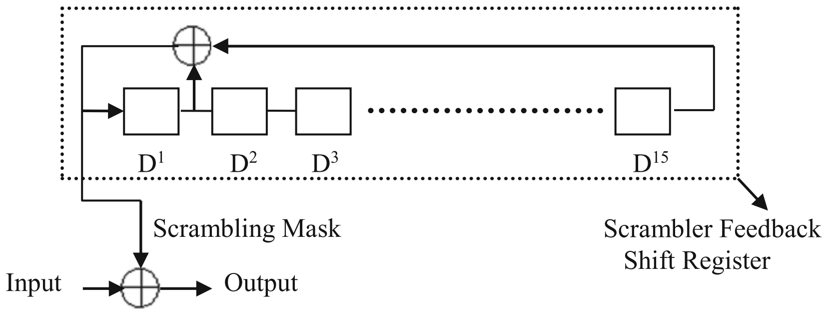


Fig. 2 Scrambler/Descrambler block diagram

3 Results and Discussions

For the simulation, parameters mentioned in Table 1, we have obtained the complementary cumulative distribution function (CCDF) graphs in MATLAB software. CCDF provides performance measurement of PAPR reduction algorithm. CCDF graphs for QPSK with phase offset of π/4 are shown in (Figs. 3, 4, 5 and 6).

In the above graph, red color indicates the CCDF for OFDM with SLM and blue color indicates the CCDF for original OFDM without SLM. Oversampling factor of four is used for accurate measurements. If this factor is less than four, then the PAPR measurement results are not accurate [1, 5, 7, 8]. The effect of QPSK phase offsets on PAPR values with SLM and without SLM has been studied as well in this article.

The results presented in Table 2 are obtained using MATLAB. From the obtained results, it is clear that the reduction in PAPR with SLM technique is significant with higher number of subcarriers (N) compared to without SLM. Also, at higher value of N, the reduction in PAPR is drastic at phase offsets of π/2 and 3π/2 compared to phase offset of π/4 even without the use of SLM technique. Thus, along with number of scrambling operations present in SLM, we have proposed that the selection of phase offsets at mapper is also very important for PAPR reduction.

In QPSK, PAPR reduction with SLM is more when compared to quadrature amplitude modulation (QAM) [9], but in this article, from Table 3, we have proposed that applying SLM at higher value of "N" is more beneficial, and we have also concluded

Table 1 Simulation parameters

Software used	MATLAB
Modulation	QPSK phase offset: (π/4, π/2, 3π/2)
No. of subcarriers (N)	64, 128, 512, 1024, and 2048
Scrambler sequence	$(+1, -1, +j, -j)$ with length "N"
Oversampling factor	4
Length of input test vector	1000

Table 2 Comparisons of PAPR values for different phase offsets and number of subcarriers

Phase offset of QPSK-modulated data	No. of subcarriers (N) (PAPR values are mentioned in dB)									
	64		128		512		1024		2048	
	With SLM	No SLM	With SLM	No SLM	With SLM	No SLM	With SLM	No SLM	With SLM	No SLM
π/4	7.12	11.34	7.57	18.55	8.13	24.12	8.86	27.13	8.89	24.31
π/2	6.87	10.48	7.46	10.8	8.50	11.29	8.86	11.85	8.94	11.97
3π/2	6.83	11.01	7.37	11.08	8.42	11.68	8.98	11.60	8.44	12.11

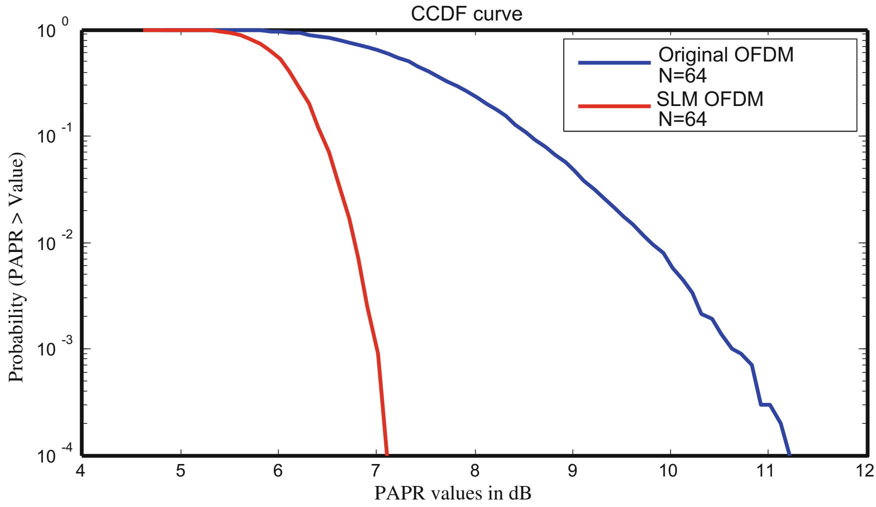


Fig. 3 PAPR for N = 64 with SLM and without SLM

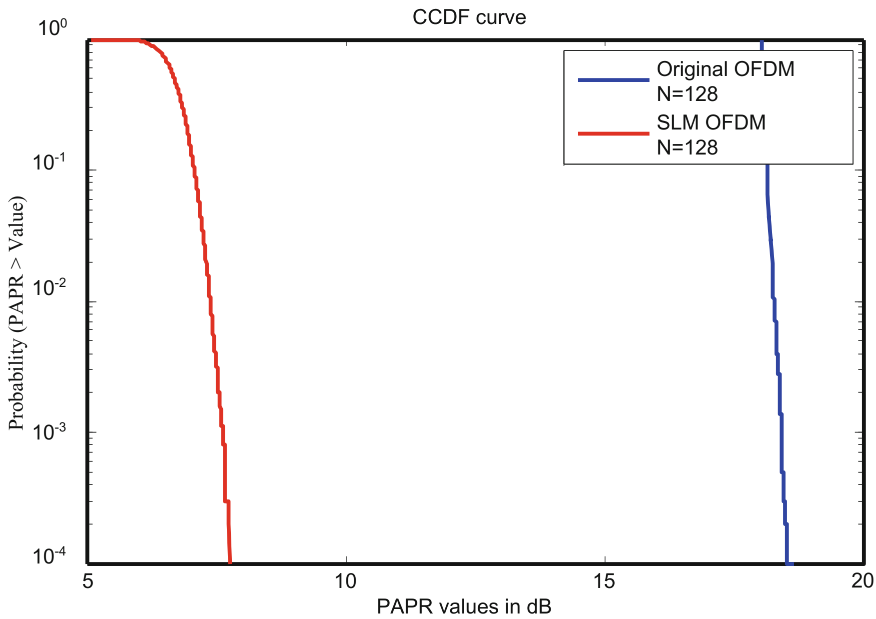


Fig. 4 PAPR for N = 128 with SLM and without SLM

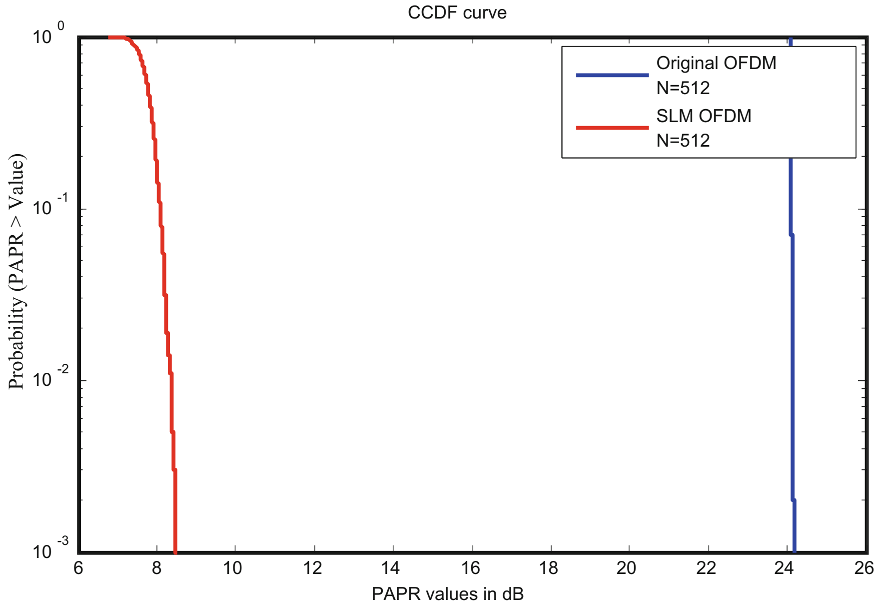


Fig. 5 PAPR for N = 512 with SLM and without SLM

Table 3 Percentage PAPR reduction with SLM at N = 64 and N = 2048

QPSK phase offset	PAPR reduction in percentage with SLM technique	
	For N = 64 subcarriers	For N = 2048 subcarriers
π/4	37.21%	63.43%

Table 4 Effect of phase offsets on PAPR without SLM technique

QPSK phase offset	PAPR reduction without SLM technique	
	PAPR For N = 64 subcarriers	For N = 2048 subcarriers
π/4	11.34 dB	24.31 dB
π/2	10.48 dB	11.97 dB
% Reduction in PAPR due to change in phase offsets and without SLM	7.58%	50.76%

from Table 4 that QPSK with phase offset of π/2 has done drastic reduction in PAPR in the absence of SLM technique. These observations were not reported in previous literatures [5, 8–10]. SLM with π/2-QPSK has provided further PAPR reduction but with increased computations. Hence, SLM phase offsets in QPSK, cost of computations, amount of PAPR reduction, power efficiency desired, and operating range of power amplifier need to be optimized based on the design requirement.

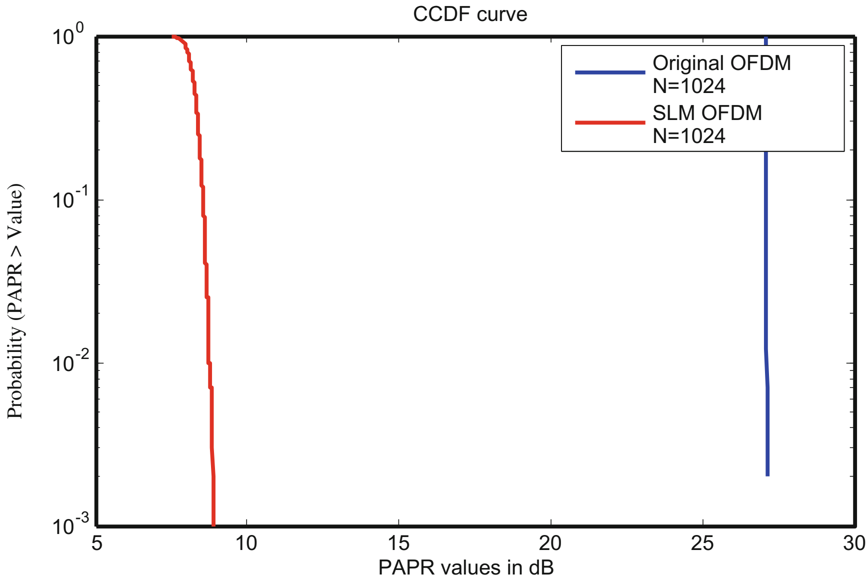


Fig. 6 PAPR for N = 1024 with SLM and without SLM

Table 5 Machine cycle for SLM scrambler implemented on Freescale SC140 architecture

Input length (in bits)	Machine cycle count	Machine cycle count/bit
1000	6370	6.37

Next, we have investigated the machine cycle requirement for SLM scrambler/descrambler algorithm using Freescale SC140 architecture. This provides an execution time requirement when SLM needs to be ported on processor, which was not presented in previous literature [9, 11–13] work (Table 5).

3.1 Computational Complexity of SLM Scrambler and Descrambler

The scrambler is working for unpacked (1 bit per byte) hard input and output. The pseudocode for implemented SLM scrambler is given below:

```
INT2 scrambler (UINT1 *u1InpPtr, UINT1 *u1OutPtr, UINT2
u2InpLen)
{
    Reg = REG_INIT //store the initial value of register
                    right aligned.
    FOR Count = 0 to InpLen DO
        XorResult = ((Reg >> SHIFTBY14) ^ Reg) & MASK_LSB
                   //compute the scrambling mask bit
        Reg = XorResult | (Reg << SHIFTBY1)
              //update register
        OutPtr [Count] = InpPtr[Count] ^ XorResult
    ENDFOR
}
```

The descrambler works on soft input and very close to scrambler in operation. The pseudocode for implemented SLM descrambler is given below:

```
INT2 descrambler (INT1 *s1Inp, INT1 *s1Out, UINT2
u2InpLen)
{
    Reg = REG_INIT
    FOR Count = 0 to InpLen DO
       XorResult = ((Reg >> SHIFTBY14) ^ Reg) &
                   MASK_LSB
                 // compute the scrambling mask bit
       Reg = XorResult | (Reg << SHIFTBY)
                 // update register

       IF XorResult THEN
   // if mask bit is 1, sign invert
          Out[Count] = -Inp[Count]
       ELSE
          Out[Count] = Inp[Count]
       ENDIF
    ENDFOR}
```

4 Conclusions

Basically, PAPR reduction using SLM occurs due to randomization of modulated data. This randomization has been performed by scrambler block present in SLM. This causes generation of polyphase sequences, thereby resulting in less PAPR. Phase offsets in QPSK also decide the amount of PAPR reduction. More scrambling results in more reduction of PAPR at the cost of increased computations. Hence, machine cycle requirement for SLM scrambler and descrambler blocks are investigated using Freescale SC140 architecture. This machine cycle profiling gives execution time requirement of SLM technique. PAPR reduction with low-computational complexity is always preferable.

References

1. B¨auml, R.W., Fischer, R.F.H., Huber, J.B.: Reducing the peak to-average power ratio of multicarrier modulation by selected mapping. Electron. Lett. **32**(22), 2056–2057 (1996)
2. Li, X., Cimini Jr., L.J.: Effects of clipping and filtering on the performance of OFDM. IEEE Commun. Lett. **2**(5), 131–133 (1998)
3. Jones, A.E., Wilkinson, T.A., Barton, S.K.: Block coding scheme for reduction of peak to mean envelope power ratio of multicarrier transmission schemes. Electron. Lett. **30**(25), 2098–2099 (1994)
4. Müller, S.H., Hüber, J.B.: OFDM with reduced peak-to-average power ratio by optimum combination of partial transmit sequences. Electron. Lett. **33**(5), 368–369 (1997)
5. Goff, S.Y.L., Al-Samahi, S.S., Khoo, B.K., Tsimenidis, C.C., Sharif, B.S.: Selected mapping without side information for PAPR reduction in OFDM. IEEE Trans. Wirel. Commun. **8**(7), 3320–3325 (2009)
6. ETSI TS 101376-5-3 v1.2.1: GEO-Mobile Radio Interface Specification, GMR-105.003
7. Breiling, M., M¨uller-Weinfurtner, S.H., Huber, J.B.: Distortionless reduction of peak power without explicit side information. In: Proceedings of the IEEE GLOBECOM'00, vol. 3, pp. 1494–1498, Nov 2000
8. Jayalath, A.D.S., Tellambura, C.: SLM and PTS peak-power reduction without explicit side information. IEEE Trans. Wirel. Commun. **4**(5), 2016–2013 (2005)
9. Dabhade, G.M., Jadhav, A.N.: PAPR Reduction using Selected Mapping in OFDM Systems. IJEEDC, ISSN (P): 2320–2084, (O) 2321-2950, COE, Bharati Vidyapeeth Deemed University, Pune, special issue-1, April 2015
10. Boonkajay, A., Adachi, F.: A blind selected mapping technique for Low-PAPR signal-carrier signal transmission. IEEE (2015). 978-1-4673-7218-3/15
11. Rapp, C.: Effects of HPA-nonlinearity on 4-DPSK/OFDM signal for a digital sound broadcasting system. In: 2nd European Conference on Satellite Communications, pp. 179–184, Oct 1991
12. Jiang, T., Wu, Y.: An: overview: peak-to-average power ratio reduction techniques for OFDM signals. IEEE Trans. Broadcast. **54**(2) (2008)
13. Al-Dalakta, E., A-Al-Dweik, A., Hazmi, A., Tsimenidis, C.: PAPR reduction using maximum cross correlation. IEEE Commun. Lett. **16**(12) (2012)

Spatial Modulation Technique: Achievements and Challenges

Namita Agarwal

Abstract Multiple antenna techniques are becoming one of the key technologies used for wireless communications these days. They trade-off higher data rates and superior error performance for increased complexity and cost. Spatial Modulation (SM) is a transmission technique using MIMO system to offer low-system complexity, improved data rate and better error performance in correlated channel environments. It exploits the properties of randomness and uniqueness of the wireless communication channel. This is done by using a coding method to establish a one is to one mapping of the transmitted information bits along with spatial positions of the transmitting antennas which are arranged in an array. The transmitted signal and the transmitting antenna number are estimated using this information for de-mapping the information block. This avoids Inter-channel Interference though a high spectral efficiency is maintained. This paper outlines the research achievements along with challenging research issues of this transmission technique.

Keywords Constellation point · Inter-channel interference · Multiple input multiple output systems · Spatial domain · Spectral efficiency

1 Introduction

Future growth of wireless communication is expected to be mainly from data-oriented services and applications. These require the throughput to increase by significant amount in comparison to the existing cellular systems. The requirement of ever-increasing data rate along with enhanced Quality of Service for next-generation networks makes it mandatory to have an improved spectral efficiency along with reliability for wireless communication. One solution to attain this is by using MIMO systems, i.e., by transmitting multiple data streams using multiple antennas. However,

N. Agarwal (✉)
Department of Electronics and Telecommunication Engineering, Don Bosco Institute of Technology, Kurla (W), Mumbai 400070, Maharashtra, India
e-mail: agarwal.namita@rediffmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_46

Table 1 Possible configurations for same spectral efficiency

<table>
<tr><th>Transmission</th><th>SM</th><th>V-Blast</th><th>Alamouti</th></tr>
<tr><td>4 b/s/hz</td><td>4×4 4QAM</td><td>2×4 4QAM</td><td>2×4 32QAM</td></tr>
<tr><td rowspan="2">6 b/s/hz</td><td>4×4 16QAM</td><td rowspan="2">2×4 8QAM</td><td rowspan="2">2×4 64QAM</td></tr>
<tr><td>2×4 32QAM</td></tr>
<tr><td rowspan="2">8 b/s/hz</td><td>8×4 32QAM</td><td rowspan="2">2×4 16QAM</td><td rowspan="2">2×4 256QAM</td></tr>
<tr><td>4×4 64QAM</td></tr>
</table>

various existing technologies using multiple antennas mostly make use of complex algorithms to eliminate the Inter-Channel Interference (ICI). Spatial modulation technique completely avoids ICI while still allowing exploitation of multiplexing gains.

Traditional modulation techniques like BPSK, QPSK, and QAM take a fixed number of bits to form a symbol which is then represented as a constellation point. In order to achieve higher spectral efficiency a new approach has been suggested which is referred as spatial modulation. In this technique, the number of transmit antenna is also used as an information source. The number of transmitting antennas and a particular signal-space diagram decide the amount of information bits which can be transmitted as shown in Table 1.

Analogous to the traditional digital modulation methods like QAM (Quadrature Amplitude Modulation), multiple antennas at the transmitter will be considered for the constellation diagram in spatial domain. Since, in (SM) [1] only one transmit antenna is active at a given moment its number can also be used as a source of information. This helps in improving the spectral efficiency of the system. Hence SM differs from the other existing MIMO techniques like space-time bit interleaved modulation where pattern of antenna is regarded as constellation but is not considered as information source [2].

A block of information bits is mapped in signal domain as a constellation point. At a particular time, only one transmitting antenna out of the total number of antennas are active and zero power is transmitted from all the other antennas. At the receiver, Maximum receive ratio combining can be used for the estimation the antenna number used for transmission along with the transmitted symbol. The spatial demodulator uses these two estimates to get back the information bits.

The content of this paper is presented in the following manner: In Sect. 2, the spatial modulation system model is presented. The advantages and disadvantages of SM have been discussed in Sect. 3. Recent research achievements have been highlighted in Sect. 4 and future scope of this field are presented in Sect. 5 and conclusions are provided in Sect. 6.

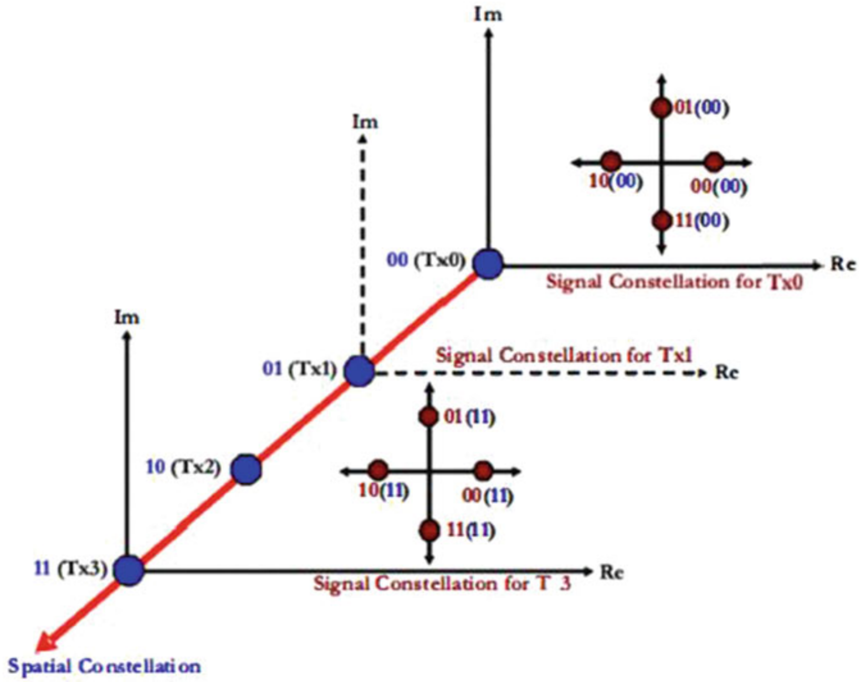


Fig. 1 Constellation diagram of SM

2 Spatial Modulation System Model

The general SM system model has wireless links with the number of transmitting and receiving antennas as N_t and N_r respectively. In SM, information bits are mapped into units carrying information [3]. Initially, a symbol is selected from a signal-space diagram, and then the index of the transmitting antenna is selected from the array. The working principle of SM is shown in Fig. 1.

A binary source generates a bit stream at the transmitter. This gets divided into blocks containing $\log_2(N_t) + \log_2(M)$ bits where $\log_2(N_t)$ and $\log_2(M)$ are the number of bits which are needed to identify the transmitting antenna and the symbol in the constellation diagram. Each block gets is split into two sub-blocks. The first sub-block helps in choosing the antenna that has to remain on for transmission and the bits of second sub-block bits are used for selecting a particular symbol present in the signal-space representation.

The signal emitted by the antenna which is active passes through a wireless channel. Each transmitted signal experiences different propagation conditions along the different transmission–reception wireless channel. Since only one transmitting antenna will be active at a particular instant, hence only one signal is received. No power is radiated by the other antennas. This random modulation introduced by the

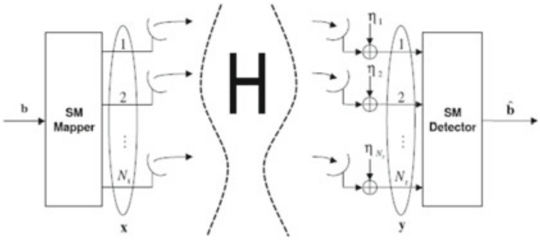


Fig. 2 Spatial modulation system model

wireless channel is used by the receiver for detection of the signal. A ML detector with complete Channel State Information (CSI) at receiver [3] is shown in Fig. 2.

In general, the number of channel impulse responses which need to be estimated is equal to $N_t N_r$. The receiver calculates the Euclidean distance between received signal and all possible signals modulated by the channel and chooses the one which is closest. In general, the number of Euclidean distances which need to be calculated is $MN_t N_r$. Likewise, all the bits of the original bit stream are recovered [3].

Thus the Spatial Modulation techniques have the following principles:

(i) Transmitted signal is naturally modulated by the wireless environment.
(ii) Each wireless link of transmission and reception has a different channel.
(iii) Receiver requires a priori knowledge of the channel for detection of the transmitted signal.

Hence, SM exploits this specific property of the wireless communication channel related to its location for communication. In this way, Spatial Modulation is different from SDMA [4].

3 Advantages and Disadvantages

This section summarizes the important advantages and disadvantages of Spatial Modulation in comparison with other MIMO techniques.

3.1 Advantages

- SM avoids ICI completely. Also, it requires only one RF chain at transmitting end as against the MIMO solutions, such as V–BLAST and Alamouti space-time schemes [4].
- A multiplexing gain which increases logarithmically as the number of transmit–antenna increase is introduced. This results in an increase by a factor of log2 (N_t). This increase is without any expansion in bandwidth of the spectral efficiency [1].

Thus spatial modulation is a highly efficient coding technique giving a code rate which is greater than one [5].
- The multiplexing gain achieved in SM helps to provide a larger capacity in comparison with other traditional less-complex coding schemes for MIMO systems, like Space-Time–Block–Codes [6].
- The receiver complexity is further reduced in SSK modulation since conventional modulation schemes are avoided [1].
- As the requirement for linearity increases for a modulation scheme, for example in QAM, there is a decrease in the efficiency of power amplifiers [7]. The results in [8] show that SM with constant-envelope modulation gives better performance than QAM.

3.2 Disadvantages

- Minimum two transmitting antennas are required to make use of SM concept.
- The transmitter to receiver wireless links need to be sufficiently different, in order to use the SM model else it may give an inadequate performance.
- For detection of data, a perfect knowledge of channel is required by the receiver. This poses constraints on the unit used for channel estimation and some overhead for channel estimation.

4 Recent Research Achievements

In recent times the research interest focuses on the applications of the Spatial Modulation as applied to MIMO wireless systems. This helps to quantify its performance compared to other popular MIMO systems. This section aims to summarize some of noteworthy results which are available.

In [1], the authors have proposed a simple receiver design based on MRC-technique to detect the transmitted bits independently. The performance analysis is done on iid Rayleigh fading channels. The comparison results with V–BLAST and Alamouti prove that SM offers has better error performance over them as shown in Table 2. A ML-optimum receiver for SM has been developed in [3]. Its comparison with the detector in [1] shows a gain in performance by about 4 dB along with a minor increase in the complexity of the receiver. The performance comparisons between V–BLAST techniques show that in Rayleigh fading channels, Spatial Modulation can offer a performance gain of about 1–3 dB as shown in Table 3.

In [9], the authors have shown results of the performance analysis between uncoded and coded systems for SSK modulation. A generalized Maximum Likelihood soft-decision decoding algorithm used in [10], shows an improvement in performance of around 3 dB in comparison to the decoding using hard decision

Table 2 Performance comparison (SNR) for 6 b/s/hz transmission (BER of 10^{-3})

	SM		V-Blast	Alamouti
	$4 \times 4$16QAM	2×4 32QAM	2×4 8QAM	2×4 64QAM
Ideal channel (dB)	18	20	21	26
Imperfect channel (dB)	26	27	30	28

Table 3 Performance comparison (Receiver complexity—6 b/s/hz transmission)

	SM		V-Blast	Alamouti
	4×4 16QAM	2×4 32QAM	2×4 8QAM	2×4 64QAM
Number of complex operations	28	14	110	15

algorithm. In [11], Trellis Coded Spatial Modulation (TCSM) concept has been used in order to reduce the effect of channel correlation.

5 Future Research Challenges

Spatial Modulation is a transmission technique which is used for data communication. Research in this field is informative years and hence fundamental issues have to be addressed to promote exploration of this technology in real propagation environments. Some open research areas in this field are summarized in this section. There is a necessity to understand the performance of SM over realistic fading conditions. Majority of the results available in the literature mainly show performance analysis in simple fading conditions and hence there is a need to explore the achievable rates over more practical scenarios to assess the full potential of the system.

The integration of Ultrawide Band with SM using in a common framework is a research area which can be explored. Further, pulse transmission methods can be used to improve the performance of the system. According to estimates published in February 2013 by Cisco [12], global mobile data traffic has grown by 70% in 2012 with 885 PB per month. The overall expected growth is 11.2 exabytes per month by 2018. Thus there is a need for wireless communications technology which is capable of meeting this forecast and also reduce the carbon footprint for next-generation cellular networks.

The deployment and optimization of cellular networks is undergoing a major shift [13, 14]. Infrastructural elements, like femto and pico BSs, are being used, which have resulted in heterogeneous cellular systems [15]. Thus the interference patterns have become more complex. The design and evaluation of communication technologies and protocols have also become more challenging [16]. Some preliminary results on evaluation of SM–MIMO communications for heterogeneous interference are shown in [17, 18].

RF energy harvesting is a concept which is being explored. It is used to provide supplementary energy supply for wireless devices [19]. The idea is to use RF signals to transport information and energy simultaneously. The objective is to charge wireless devices with the help of microwave radiation links [20, 21].

6 Conclusion

This paper summarizes the principle, advantages, disadvantages, recent research achievements and research issues of Spatial Modulation which is a low complexity MIMO transmission technology. It is a physical layer transmission technique. This novel technique is a combination of digital modulation, coding, and multiple-antenna transmission. It utilizes the location-specific property of the transmission channel used for communication. Here, the location of every transmit-antenna in the array is used for conveying information. It is a power efficient transmission scheme which provides better energy efficiency than conventional MIMO schemes.

Recent research achievements and results discussed in this paper indicate that SM is a promising candidate which can be used for low-complexity MIMO system implementation. However, research applications in this field are still in nascent stage and hence many issues will have to be addressed to understand the potential and limitations of this technology in more realistic environments of propagation.

References

1. Mesleh, R., Haas, H., Sinanovic, S., Ahn, C.W., Yun, S.: Spatial modulation. IEEE Trans. Veh. Technol. **57**(4), 2228–2241 (2008)
2. Tonello, A.: Space-time bit-interleaved coded modulation with an iterative decoding strategy. In: Proceedings of the IEEE 52nd Vehicular Technology Conference vol. 1, pp. 473–478 (2000)
3. Jeganathan, J., Ghrayeb, A.: LSzczecinski: spatial modulation: optimal detection and performance analysis. IEEE Commun. Lett. **12**(8), 545–547 (2008)
4. Mietzner, J., Schober, R., Lampe, L., Gerstacker, W.H., Hoher, P.A.: Multiple–antenna techniques for wireless communications—a comprehensive literature survey. IEEE Commun. Surv. Tutor. **11**(2), 87–105 (2009)
5. Song, S., Yang, Y., Xiong, Q., Xie, K., Jeong, B.J., Jiao, B.: A channel hopping technique I: theoretical studies on band efficiency and capacity. In: IEEE International Conference on Communication, Circuits and Systems, vol. 1, pp. 229–233 (2004)
6. Yang, Y., Jiao, B.: Information–guided channel–hopping for high data rate wireless communication. IEEE Commun. Lett. **12**(4), 225–227 (2008)
7. Hasan, Z., Boostanimehr, H., Bhargava, V.K.: Green cellular networks: a survey, some research issues, and challenges. IEEE Commun. Surv. Tutor. **13**(4), 524–540 (2011)
8. Di Renzo, M., Haas, H.: Bit error probability of SM-MIMO over generalized fading channels. IEEE Trans. Veh. Technol. **61**(3), 1124–1144 (2012)

9. Jeganathan, J., Ghrayeb, A., Szczecinski, L., Ceron, A.: Space shift keying modulation for MIMO channels. IEEE Trans. Wirel. Commun. **8**(7), 3692–3703 (2009)
10. Hwang, S.U., Jeon, S., Lee, S., Seo, J.: Soft–output ML detector for spatial modulation OFDM systems. IEICE Electron. Express **6**(19), 1426–1431 (2009)
11. Mesleh, R.Y., Di Renzo, M., Haas, H., Grant, P.M.: Trellis coded spatial modulation. IEEE Trans. Wirel. Commun. **9**(7), 2349–2361 (2010)
12. Cisco: Cisco visual networking index: global mobile data traffic forecast update, 2012–2017 (2013)
13. Di Renzo, C., Merola, A., Guidotti, F., Santucci, G.E.: Corazza: error performance of multi-antenna receivers in a Poisson field of interferers VA stochastic geometry approach. IEEE Trans. Commun. **61**(5), 2025–2047 (2013)
14. Di Renzo, M., Guidotti, A., Corazza, G.E.: Average rate of downlink heterogeneous cellular networks over generalized fading channels VA stochastic geometry approach. IEEE Trans. Commun. **61**(7), 3050–3071 (2013)
15. Ghosh, A., Andrews, J.G., Mangalvedhe, N., Ratasuk, R., Mondal, B., Cudak, M., Vi-sotsky, E., Thomas, T.A., Xia, P., Jo, H.S., Dhillon, H.S., Novlan, T.D.: Heterogeneous cellular networks: from theory to practice. IEEE Commun. Mag. **50**(6), 54–64 (2012)
16. Andrews, J.G.: Seven ways that HetNets are a cellular paradigm shift. IEEE Commun. Mag. **51**(3), 136–144 (2013)
17. Di Renzo, M., Haas, H., Ghrayeb, A., Sugiura, S., Hanzo, L.: Spatial modulation for Generalized MIMO: challenges, opportunities, and implementation. In: Proceedings of the IEEE, vol. 102 (1), pp. 56–103 (2014)
18. Lu, W., Di Renzo, M.: Performance analysis of spatial modulation MIMO in a poisson field of interferers. In: Proceedings of the IEEE International Conference Computing, Networking and Communications, pp. 1–7 (2014)
19. Sudevalayam, S., Kulkarni, P.: Energy harvesting sensor nodes: survey and implications. IEEE Commun. Surv. Tutor. **13**(3), 443–461 (2011)
20. Huang, K., Lau, V.K.N.: Enabling wireless power transfer in cellular networks: architecture, modeling, and deployment (2012). http://arxiv.org/pdf/ 1207.5640.pdf
21. Zhang, R., Ho, C.K.: MIMO broadcasting for simultaneous wireless information and power transfer (2013). http://arxiv.org/pdf/1105.4999.pdf

A Novel Cluster Based Algorithm for Outlier Detection

Manish Mahajan, Santosh Kumar and Bhasker Pant

Abstract Nowadays an important issue as well as challenge in data mining is obviously is outlier detection. Outlier detection has been used in many areas such as Fraud detection, Intrusion detection, Health care, Fault detection, etc., where detection of outliers is based on the different characteristics of data or datasets. In this current age of 'Information Technology', large numbers of processes are obtainable in the domain of data mining to discover the outliers by successfully creating the clusters and after that detecting the outliers from these created clusters. In data mining, cluster methods are highly essential and have been applied from micro- to macro-applications. Basically clusters are a pool of similar data objects put together grounded on the attributes and district features they have. Specifically outlier detection is used to recognize and exclude inconsistency from the available data sets. In the presented work an algorithm has been suggested which is based on clustering approach to the given data sets. The proposed algorithm efficiently detects outliers inside the clusters by using clustering algorithm and weight based approach.

Keywords Data mining · Outlier · Outlier detection · K-means clustering

1 Introduction

As day by day the digital world is growing so fast, there is a tremendous growth of computer hardware and software with a resulting increase in the dependency of business on data. From small business to large business houses, they all work on the information superhighway and generate and consume huge amounts of data in

M. Mahajan · S. Kumar (✉) · B. Pant
Graphic Era Deemed to be University, Dehradun, India
e-mail: amu.santosh@gmail.com

M. Mahajan
e-mail: manish.mhajn@gmail.com

B. Pant
e-mail: pantbhaskar2@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_47

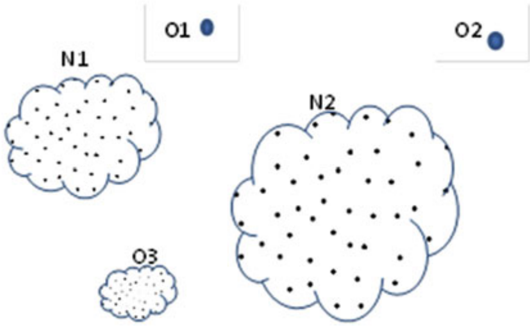


Fig. 1 Outliers N1 and N2 are two given normal data. O1 and O2 are two outliers and points in O3 are observed as outliers

their routine operations. As a result, huge quantity of data has been collected and stored in databases [1, 2]. Researchers working in the field have projected that, the quantity of information being produced by business globally doubles every 2 years. To automate the processes of data dissemination and analysis for these small or big business houses data mining is required. These techniques extract the relevant information and provide the results to the organizations. The result thus obtained is highly valuable to the organizations for the process of decision making [3–5]. The work handed over to a pool of data objects is regarded as clusters, these are kept in a pool so that each and every element in a cluster are alike to each other than to those in other clusters, and this process of clustering is cumulatively called as cluster analysis [6]. Outlier is an element whose values are dissimilar to the all the other presented elements in that particular group. This makes the outlier detection and analysis an essential application in mining to find out the nonstandard points for various applications, for instance intrusion detection, network sensors, fraud detection, stock market study, and advertising.

The fundamental idea to identify an outlier is to first of all to identifying anomalous points between the large numbers of available data points. Outlier revealing is a significant research problem with the ultimate objective to recognize all the elements that are slightly or totally different from others in the given database [7–9] as shown in Fig. 1. Outliers can be originated in diverse domains like the dataset in the field of medical science and in social networking. Effective mean of detection of an outlier certainly decreases the likelihood of making such conclusions which can hamper the success of any domain [10, 11].

Outlier detection finds abnormal patterns in any dataset that do not fit in with expected behavior. These non-adjusting patterns are frequently referred to as anomalies, novelties, outliers, or, exceptions. Anomaly and Outliers are the two most commonly used terms. Another definition of an outlier states that an outlier is a data object which is deviated from the other data object and creates suspicious environment. It is highly essential to detect an anomaly as it causes data interpret to substantial information in diverse diversity of applications [12, 13].

Numerous algorithms of data mining and machine learning, for arithmetical operation and statistical operation sometimes cannot work well in the occurrence of outliers. Therefore there is an acute need of a novel outlier detection algorithm by using enhanced k-means clustering so that the outliers can be removed effectively [14, 15].

2 State of Art

There are large numbers of popular state of art available in the field of outlier detection; some of the recent to fulfill the aim of the proposed work will be discussed in the given section. There are mainly three sub-categories of outlier detection, and all these categories are highly essential and recommended in identifying and removing the outliers.

1. **Distance-based outlier detection**: distance-based outlier detection identifies the outlier from the nearest points while considering the distance factor.
2. **Density-based outlier detection**: density-based outlier detection identifies the outliers from the nearest point considering the density issue.
3. **Distribution-based outlier detection**: distribution-based outlier detection identifies the outliers by using specific arithmetical model.

Purohit [16] proposed an enhancement over the traditional k-means clustering algorithm to overcome some of its limitations. One of the major reasons why the traditional k-means algorithm performs poorly is due to the fact that the initial centroid points are selected randomly. The proposed algorithm improves the performance and quality of clusters for the traditional k-means algorithm by addressing this limitation of centroid selection by selecting the k initial cluster centroids in an efficient manner instead of doing it randomly. The proposed algorithm calculates Euclidian distance between the various data points in order to identify the closest data objects. The proposed algorithm provided more accurate results and also helped to reduce the mean square distance. But the proposed algorithm failed to perform effectively when applied to a sparse data set. Shunye [17] proposed an enhanced k-means clustering algorithm consisting of three main steps. The first step works towards the construction of a dissimilarity matrix. The dissimilarity matrix is used to generate a Huffman tree by applying the Huffman algorithm in the second step. The initial k centroids are obtained from the output of the Huffman tree. In the third step, the k-means clustering algorithm is used with the initial centroids to obtain the k clusters. The proposed algorithm was tested using the Wine and Iris datasets from the UCI Machine learning repository. The algorithm provided better results and curacy rates as compared to the traditional k-means clustering algorithms. Fahim [18] suggested an efficient k-means algorithm to overcome the limitations of the traditional k-means algorithm. Although the traditional k-means algorithm is known for its ease of operation, straightforwardness and its efficiency even with sparse data; it still faces some limitations. One of the major limitations is the dependence of final results on the selection of initial centroids. The proposed algorithm initially allocates datasets to their nearby centroids and then calculates their distance from other centroids. It uses two distance functions: a distance() function initially measure distance between the dataset and the nearest cluster head, and a distance_new() function to find the distance between the data objects and the other remaining clusters. These distance functions are used to decide whether to keep the dataset with the initially decided centroid or to move it to a new one. Thus, the proposed algorithm works towards reducing the time and increasing the efficiency of the traditional k-means algorithm. The experimental results prove the speed and efficiency of the proposed algorithm over the traditional k-means algorithm.

Wang [19] proposed an improved k-means algorithm to deal with the limitations of outlier detection using traditional k-means clustering algorithm. The proposed algorithm used the noise data filters to handle the outliers in the datasets being tested. Density-based outlier detection methods were used to identify the outlier data points or outlier datasets and these were subsequently removed. This ensured that the outlier data points are not involved while computing the initial cluster centers. The proposed algorithm used clustering time and accuracy to check the efficiency. This algorithm worked effectively with smaller to medium data sets, but when applied to large data sets, the algorithm tended to take quiet long to produce results.

Mahmud et al. [20] proposed an algorithm using a heuristic approach to generate the k initial centroids. The proposed algorithm performed better in terms of both accuracy and computational time. The proposed algorithm analyzed all the multi-dimensional data objects having weight factors. The algorithm then used Merge-sort to arrange the output thus obtained. The data points are then arranged into k clusters and the data point nearest to the possible mean is taken as the initial centroid. Chauhan et al. [21] discussed the many approaches of anomaly detection done by using K-Means clustering algorithm for dataset in combination with certain other methods. They also discussed the different areas where outlier or anomaly detection is employed.

3 Methodology

In the proposed approach the outliers can be detected in two phases. In the First phase we used improved clustering algorithm to provide a set of clusters. In the second phase for each cluster the absolute space among every object and the cluster center is estimated using proposed weight-based center approach (Fig. 2).

3.1 Input Data Set

The dataset can be collected from various repositories like UCI etc.

Fig. 2 Proposed methodology

3.2 *Clustering the Dataset*

It is a popular method successfully implemented to clutch the identical data points in the given pool or clusters. Clustering is regarded as an essential tool for the analysis of outlier.

3.3 *Outlier Detection Using Weight Based Center Approach*

A tremendously significant job in a range of various recent upcoming fields is outlier detection. Outlier detection is basically a method or a process to accurately find out the objects that are different or unpredictable concerning to the rest of the data or which are in remote distance from centroids of the cluster. First calculate the weight of each cluster using Eq. (1) and then store the result into a Vector W.

$$W_k = \sum_{x=1}^{n} d_x \tag{1}$$

where k is the cluster number (k = 1, 2,…, n) and dx is the number of data items (d1, d2,…, dn) in particular cluster K. Next, find the mean of each cluster using Eq. (2) and then the result stored into a Vector M

$$M_k = \frac{\sum_1^n d_x}{T_n}, \tag{2}$$

where k is the cluster number (k = 1, 2,…, n) and dx is the number of data items in particular cluster k and Tn is the total number of data items in cluster K. Calculate the Maximum and Minimum value of each cluster and then store the maximum value into an Array Dk-max and store the minimum value in Dk-min Array. Dk-max = Max (d1, d2,…, dn) Dk-min = Min (d1, d2,…,dn) where k = 1, 2…, n is the cluster number. The threshold value for each cluster K(1, 2,…, n) can be calculated and then the results stored into an array Th using following Eq. (3)

$$Th_{k-critical} = |(M_k - (D_{k-max} + D_{k-min})/2)|, \tag{3}$$

where k is the cluster number. Compare each data item dx(x = 1, 2,…, n) in a particular cluster K (1, 2,…, n) with the threshold value Thk-critical. If the value is found less than Thk-critical then the data item is the outlier.

4 Proposed Algorithm

Phase 1: Cluster Detection

Data Set D = {d1, d2,…, dn} for n data points

Cluster Center C = {c1, c2,…, ck) ci is the center of cluster i and k is the number of clusters

- **Step 1**: Start
- **Step 2**: Commence the group of clusters, S, as the blank set, read a new object p.
- **Step3**: Generate a cluster with the object p.
- **Step 4**: If no objects left in then go to step 6, then read a new object p, and find the
- cluster C* in S that is closest to the object p. It is also described that, find a cluster C* in S, such that for all C^- in S, d(p, C*) ≤ d(p, C^-).
- **Step 5**: If d(p, C *) > r, go to step 2.
- **Step 6**: Combine object p into cluster C* and amend the Cluster Summary Information (CSI) of cluster C*, go to step 3.
- **Step 7**: End

Phase 2: Outlier Detection

- **Step1**: Calculate the weight-based center Wk as given in the Eq. (1).
- **Step 2**: Calculate the Mean of each cluster center, let it be Mk as given in the Eq. (2).

- **Step 3**: Calculate the maximum and minimum value of each cluster K, where Dk-max = Max (d1,…, dn) and Dk-min = Min(d1, d2,…dn).
- **Step 4**: Calculate Thk-critical = ABS (Mk − ((Dk-max + Dk-min)/2) using Eq. (3).
- **Step 5**: Compare each data item dx in a cluster with the threshold value Thk-critical where k is the cluster number.
- **Step 6:** If the data value is dx found less than Thk-critical then the given data item in cluster k is the outlier, where x = 1, 2,…n.
- **Step 7:** Remove the outlier data items from the cluster.
- **Step 8**: Repeat step 8–13 for each resultant cluster.

5 Conclusion

In this IT era, detection of outlier is considered as data mining's primary task. Discovering outliers is the task that finds a large number of dissimilar data objects with respect to the entire set of data. The proposed algorithm efficiently detects outliers inside the clusters by using clustering algorithm and weight-based approach. In this work, we first group the data items into pool of clusters based on similarities between them. The computation time reduced considerably because of reduction in size of dataset. After that the outlier can be detected in each cluster using threshold value that can be calculated programmatically. In future, we prove the results experimentally and improvements can also be done in proposed approach.

References

1. Cateni, S., Colla, V., Vannucci, M.: Outlier detection methods for industrial applications. Advances in robotics. In: Automation and Control, pp. 274–275 (2008)
2. Ahmad, A., Dey, L.: A k-mean clustering algorithm for mixed numeric and categorical data. Data Knowl. Eng. **63**, 502–527 (2007)
3. Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. Artif. Intell. Rev. **22**(2), 85–126 (2004)
4. Fawzy, A., Mokhtar, H.M.O., Hegazy, O.: Outliers detection and classification in wireless sensor networks. Egypt. Inf. J. **14**, 157–164 (2013)
5. Khan, F.: An initial seed selection algorithm for k-means clustering of geo-referenced data to improve replicability of cluster assignments for mapping application. Appl. Soft Comput. **12**, 3698–3700 (2012)
6. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3), 1–58 (2009)
7. Pachgade, S.D., Dhande, S.S.: Outlier detection over data set using cluster-based and distance based approach. Int. J. Adv. Res. Comput. Sci. Soft. Eng. **2**(6), 12–16 (2012)
8. Zhu, C., Kitagawa, H., Papadimitriou, S., Faloutsos, C.: Outlier detection by example. J. Intell. Inf. Syst. **36**, 217–247 (2011)
9. Shi, Y., Zhang, L.: COID: a cluster–outlier iterative detection approach to multi-dimensional data analysis. Knowl. Inf. Syst. **28**, 710–733 (2010)
10. Indira Priya, P., Ghosh, D.K.: A survey on different clustering algorithms in data mining techniques. Int. J. Mod. Eng. Res. **3**(1), 267–274 (2013)

11. Gupta, M., Gao, J., Aggarwal, C.C., Han, J.: Outlier detection for temporal data. In: Proceedings of the 13th SIAM International Conference on Data Mining (SDM) (2013)
12. Divya, T., Christopher, T.: A study of clustering based algorithm for outlier detection in data streams. Int. J. Adv. Netw. Appl. (IJANA) (2015). ISSN 0975-0282
13. Chugh, N., Chugh, M., Agarwal, A.: Outlier detection in streaming data a research perspective. Int. J. Sci. Eng. Technol. Res. (IJSETR) **4**(3) (2015)
14. Bhosale, S.V., et al.: Outlier detection in straming data using clustering approached. Int. J. Comput. Sci. Inf. Technol. (IJCSIT) **5**(5), 6050–6053 (2014)
15. Manoharan, J.J., Hari Ganesh, S.: Improved k-means clustering algorithm using linear data structure list to enhance the efficiency. Int. J. Appl. Eng. Res. **10**(20) (2015). ISSN 0973-4562
16. Purohit, P.: A new efficient approach towards k-means clustering algorithm. Int. J. Comput. Appl. **65**(11) (2013)
17. Shunye, W.: An improved k-means clustering algorithm based on dissimilarity. In: 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC) Dec 20–22, 2013, Shenyang, China. IEEE
18. Fahim, S.A.M., Torkey, F.A., Ramadan, M.A.: An efficient enhanced k-means clustering algorithm. J. Zhejiang Univ. Sci. A. ISSN 1009-3095, ISSN 1862-1775
19. Wang, J., Su, X.: An improved k-means clustering algorithm. IEEE (2011)
20. Mahmud, Md.S., Rahman, Md.M., Akhtar, Md.N.: Improvement of k-means clustering algorithm with better initial centroids based on weighted average. In: 2012 7th International Conference on Electrical and Computer Engineering, 20–22 Dec 2012, Dhaka, Bangladesh. IEEE (2012)
21. Chauhan, P., Shukla, M.: A review on outlier detection techniques on data stream by using different approaches of KMeans algorithm. In: 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA), IMS Engineering College, Ghaziabad, India. IEEE (2015)

Sentimental Analysis of Twitter Data on Hadoop

Jayanta Choudhury, Chetan Pandey and Anuj Saxena

Abstract Data is something without which organizations can never reach any conclusion and cannot extract any particular pattern. These data sets are the sources on which organizations rely while taking important strategic decisions. There are many social platforms on which people around the world are accessing and these platforms are generating a huge amount of data. This data can be differentiated on the basis of their volume, velocity and variety. Organizations term such a huge amount of data as Big Data. These social data sets are of great use for improving business strategies. Nowadays, twitter has become a great social platform for expressing different opinions. This paper focuses on MapReduce-based sentiment analysis of data received through twitter. The data is first cleaned to retain only text, then MapReduce is applied to get the frequency of each word which is then matched with the dictionary created for positive and negative words over Hadoop environment. The results are compared with Naïve Bayes and SVM classifier. It has been observed that time consumed by the proposed system is 45% less than SVM and 38% less than Naïve Bayes. The accuracy in terms of a total number of words detected, positive and negative words, was also observed to be 11%, 16%, 18% respectively in case of SVM and 9%, 13%, 16% respectively in case of Naïve Bayes.

Keywords Big data · Hadoop · MapReduce · SVM · Naïve Bayes

J. Choudhury (✉) · C. Pandey
Graphic Era University, Dehradun, India
e-mail: jayanta34chou@gmail.com

C. Pandey
e-mail: schetanpandey@gmail.com

A. Saxena
Institute de informatica, Dehradun, India
e-mail: ididoon@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_48

1 Introduction

For past few years, researchers and scholars have made Hadoop MapReduce [1, 2], a clear choice for processing a large data set consisting of unstructured data in it. MapReduce is a distributed parallel programming model to process a large volume of data in a homogeneous environment. When a client sends a data set to MapReduce, it splits the data into small chunks called Splits and organizes them into key-value pairs for parallel processing. The Mapper function is then applied on each split. It takes the key-value pairs of the splits and generates a list of key-value pairs. Then, shuffling is done. In shuffling, it sorts the data according to their values. The Reducer function is then applied and it takes the sorted key-value pairs and generates an aggregate and generates new key-value pairs as output. This procedure allows parallel processing of large data set and generates the output within a minimal time. Social Media plays an important role in guiding public opinion over many strategic issues. A fast robust technique is required as the data through Social Media is huge, accuracy is generally compromised with the increase in the volume of data. As the data dictionary of positive, negative words is large and the data collected even for seven days from Twitter is huge, this paper performs Sentimental Analysis on Twitter data using Hadoop MapReduce framework to reduce the time consumed and improve the accuracy. After initial filtering of the data on the fly the data is passed through Mapreduce which reduces the time consumed considerably of the preprocessing and matching with the dictionary as, in a conventional method the program reads a word from the dictionary first and then tries to find match in the tweets collected this requires lot of iterations and resultantly lot of time. The proposed method creates a list of words with their recurrence in the tweet data and matched with the dictionary, a match would give its occurrence immediately.

2 Motivations

In this paper, beginning with an introduction of the working procedure of Hadoop MapReduce with a Word Count example and then demonstrate the Sentimental Analysis of Twitter data using Hadoop MapReduce framework.

2.1 Motivating Example

Most basic example of MapReduce is the Word Count problem. At the start when working on Hadoop MapReduce, scholars start with this example to understand the working of MapReduce. In MapReduce there are four stages Splitting, Mapping, Shuffling, and Reduce. When a dataset is passed to the MapReduce, it first does the Splitting stage where it logically partitions the dataset into small chunks called Splits. It partitions the dataset into Splits and represents them as key-value pairs. After the splitting is done, the Mapping function then applied to each split. It consumes the

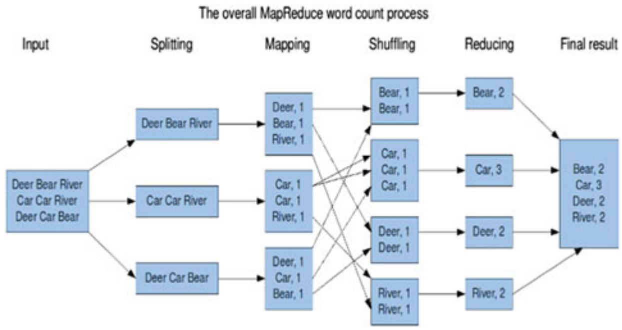


Fig. 1 Word count using MapReduce

key-value pair of each split and generates a list of new key-value pairs. After the Mapping stage is over then those newly generated key-value pairs in the Mapping stage are sorted in the Shuffling stage. The key-value pairs are sorted according to the values and passed to the Reduce stage. In the Reduce stage those sorted key-value pairs are consumed and generate an aggregate and finally generates new key-value pairs as output (Fig. 1).

Taken from https://cs.calvin.edu/courses/cs/374/exercises/12/lab/.

3 Related Work

In the paper "Sentimental Analysis of Social Media Using R Language and Hadoop: Rhadoop" [3], authors have performed sentimental analysis of twitter data using R language. It collects the sentiment information in the form of either positive scores negative scores or in between them. Twitter data whose size is in TBs was analyzed using R language and Rhadoop connector. This paper also focused on the performance estimation on two different platforms R language and Rhadoop tool. The R language is powerful and suitable for implementation of data extraction and data analysis tool. When the size of data exceeds the size of physical memory of R environment then sometimes R gives poor results and also terminates the R session. To solve this problem Rhadoop was used. In the paper [4], the author has described the use of a domain-specific dictionary for improving the accuracy of sentiment analysis in a particular domain. Author has compared four dictionaries representing a distinct dictionary building methods for identifying which method have higher or lower potential in building a domain dictionary for sentiment analysis in software engineering texts. Author has described procedure for improving the accuracies and building a domain-specific dictionary for improving the performance of sentiment analysis in software engineering texts.

4 Solution for Sentimental Analysis

4.1 Data Preprocessing

Data preprocessing is an important step in the analysis. Real-world data consists of missing values, noisy data, lacks in certain attributes of interest and sometimes not logically correct. In this paper data preprocessing has been achieved at the beginning of the sentiment analysis. The steps are as follows:

1. Data cleaning is the first step in data preprocessing. In live twitter dataset there are many unwanted characters like "\t", "\n" and many more. In this paper, all unwanted characters were eliminated using delimiters in the mapper().
2. Data compression is the most vital step to handle big data. In this paper, data compression has been achieved after the completion of the Reduce stage. In MapReduce the reduce() stage produce the final set of output based on the key.
3. Data normalization is also important because it either sorts the data in a canonical form or in the normal form. In this paper, the data normalization has been achieved at the shuffling stage. In this stage, the mapped object produced by the mapper() are sorted based on their keys.
4. In this paper, data transformation is achieved in the Map phase. The input data from the dataset is transformed into key/value pairs.
5. Feature extraction has been achieved in the Reduce phase. In the Reduce phase all redundant data were reduced.
6. Feature selection has been achieved in the shuffling stage as it sorts the mapped objects and provides the reducer which makes it easy for the reducer to produce the final output.

4.2 Sentiment Analysis Using MapReduce

Sentimental Analysis is attracting various scholars from all over the world. Some top MNCs like Facebook, Twitter and others are also focusing on analyzing the sentiments of their users. In this paper, the focus will be on Twitter data. Twitter data are used on different platforms like political analysis, hospitality industries and many others. To perform sentimental analysis there are mainly three steps. They are as follows:

1. Extraction and Storing Live Twitter Data: Extraction of the live twitter data is only possible after getting authorized keys. To get authorized keys firstly, create a twitter account. Secondly, create a twitter application by going to the Twitter Development services. While creating an application on Twitter, it will provide Authorized Consumer key, Consumer Secret key, Authorized Token key and Secret Token key. Use theses keys to extract live twitter data and store these data into a file. In this paper, twitter data has been stored in a text file.
2. Extracting the useful data from those stored twitter data: The stored twitter data consists of tab separated data, space separated data and many more. So, use

delimiters in the code to eliminate "\n", "\t" and others. Use MapReduce to get the words and phrases from the stored twitter data.

3. Matching those useful data with the dictionary for predicting the positive or negative sentiments: After getting the words and phrases from the stored twitter data, match those with positive words and negative words. While matching the words count of positive matches and negative matches. Finally display the sentiment count.

4.3 Matching the Data with the Dictionary

The sorted output of the twitter dataset by MapReduce is stored into a HashMap array as the output consists of the words and its counts. It becomes easy to compare the twitter data from the HashMap array with the dictionary. In this paper, two dictionaries have been used one containing the positive words and the other contains all the negative words. A HashMap array and compares individually with the positive words and negative words dictionary. After the comparison, the number of positive and negative matched words is again compared to each other. If the number of positive words is more than the result will be displayed as positive sentiment or else it will display negative sentiment as a result.

5 Experiment Setup

This experiment was carried out on a single node Hadoop-2.7.4 cluster. The cluster consists of one master computing node and one slave node. Compute node has the physical environment, i.e., one Intel CORE i5 CPU, with 8 GB memory, Windows 10 OS. The Sentimental Analysis has been used to find out the positive or negative sentiments from the gathered live Twitter data by comparing the words from the Twitter data with a positive words dictionary and a negative words dictionary. The experiment dataset, consists of various tweets on Gujarat elections, GST, Demonetization and Cast Reservation Quota. This dataset is being used for finding the sentiment of the common public about the present situation (Table 1).

Table 1 Project information table

Data	Twitter data
Hash tags	#gujaratelections, #gst, #demonetization, #reservation
Duration	7 days
Size	1 MB–1 TB (different data packet sizes)

5.1 Steps Followed

Step1: Create a Twitter account.
Step2: Go to the TwitterDev site for creating an application.
Step3: After the twitter application is created authorized keys will be available.
Step4: Create a code for live twitter data extraction and use those authorized keys in the code for data extraction. Then save those extracted data in a file in.txt format.
Step5: Perform data preprocessing on the twitter data with MapReduce.
Step6: Store the dictionary and preprocessed data in arrays.
Step7: Match the preprocessed data with the dictionary.
Step8: Then compare the number of positive and negative word match.
Step9: Display the result.

5.2 Algorithm

```
Algorithm Collect Twitter data
Setup:
Create Authorization key and Authenticate Twitter
Working:
Step 1. query ← collect tweet based on #tag
Step 2.  Iterate till size of query
             Remove special characters using regular expressions from query
             Tokenize query and write to file
         end iteration
Pre-Process Data further using Map
Step3  Open file and iterate until end of file
Step4.       If data exist then set the input value type and the type of output key
Step5.           Iterate till more tokens
                     Map←       set word as each input keyword
                                Collect output as (word, one)
                                Emit to reducer based on primary key
Step6.           end iteration
Pre-Process Data further using Shuffle:
Step7. Iterate until end of file
             store line number as key and data as the value
             sort file with key and value store as Mapper output
Step8. end iteration
Pre-Process further using Reducer
Step9. Iterate until end of file
             sum similar words store keyword and sum as an output
Step10. end iteration
Loading dictionaries and the pre-processed data
Step11. Store the pre-processed data← create a HashMap array
```

Step12. Iterate through data
Tokenize and put the data in HasMap array
Step13. end iteration
Step14. Positive_words dictionary← scan the file
Create an arraylist to store Positive_words dictionary
Step15. while file has records
Add records from the file to the arraylist
Step16. end loop
Step17. Negative_words dictionary← scan the file
Create an arraylist to store Negative_words dictionary
Step18. Iterate till file has records
Add records from the file to the arraylist
Step19. end arraylist
Analysis
Step20. Iterate till hashmap is not empty
If record matches with Positive_words dictionary
Print and count the word matched
Step21. end iteration
Step22. Repeat steps 20-22 for Negative_words
Step23. end loop
Step24. Compare negative and positive word count and print appropriate result.

6 Results and Comparison

For experimental purpose, the initial dataset implemented was of size 1,232 KB. The dataset consists of live Twitter data collected by running a Twitter data search code (Figs. 2 and 3).


null
iPhone8?????…??????null
iPhone8?????…??????
RT @SugarHairy: Cuando quieres el iPhone8 ?
https://t.co/Z7gSxAamDtnull
iPhone8?????…??????
RT @SugarHairy: Cuando quieres el iPhone8 ?
https://t.co/Z7gSxAamDt
A?iPhone8?????
B???????
A??????????????????????? https://t.co/Iw7MIAld8ynull
iPhone8?????…??????
RT @SugarHairy: Cuando quieres el iPhone8 ?
https://t.co/Z7gSxAamDt
A?iPhone8?????
B???????
A??????????????????????? https://t.co/Iw7MIAld8y
RT @N2xpcymZLkRahYM: ?????????
iPhone8?????1????3????

??????????&RT
?????????UP!

???????????

?????&RT?????????DM?????????????

#???????…null
iPhone8?????…??????

Fig. 2 Twitter data

Fig. 3 Preprocess Twiter data

```
"you  291
#15   484
#??????     680
#?????????????????good      195
#HiGH_ï¿½  582
#HiGH_ï¿½null    98
#MersalDiwali    974
#TereMereReprise 1371
&     877
(?????+3000)     779
(INT  1171
,     388
11/3  1171
2K    1171
5??   779
6000?~10000?     779
?     974
??    1558
??&?? 779
??:   779
???   2532
?????!     779
?????check?????????????     680
?????check?????????harder"  1
```

Using MapReduce Word Count program the data in the dataset has been counted and separated. It shows the words and count of the particular word. After the words are separated, taking each word and comparing the word with the dictionaries (Fig. 4).

Figure 5 showing the comparison of a number of words detected when sentiment analysis per formed on this system, SVM and Naïve Bayse. Figure 6 showing a comparison of a number of positive words detected when sentiment analysis performed on this system, SVM and Naïve Bayse. Figure 7 showing a comparison of a number of negative words detected when sentiment analysis per formed on this system, SVM and Naïve Bayse. Figure 8 is showing the comparison of time unit in seconds to complete the whole job when sentiment analysis performed on this system, SVM and Naïve Bayse. Figure 9 shows the time taken to perform sentiment analysis on different data packet sizes. The starting data packet size was of 1 MB and gradually been increased to test the performance. For the final testing procedure, the data packet size was of 1 TB. As a result of performing sentiment analysis on different sizes of the data packet, performance could be measured very neatly (Tables 2 and 3).

```
                Reduce input groups=195
                Reduce shuffle bytes=2902
                Reduce input records=195
                Reduce output records=195
                Spilled Records=390
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=22
                Total committed heap usage (bytes)=379584512
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1220608
        File Output Format Counters
                Bytes Written=2571
total time taken by this program in milli second is : 7990
total time taken by this program in second is : 7.99
this is before file1
this is before file2
the positive words from the input are the following::
positive_word_is: darling
positive_word_is: enough
positive_word_is: good
positive_word_is: like
positive_word_is: right
total_positive::5
negative words from the input are the following::
negetive_word_is: harassment
negetive_word_is: hate
negetive_word_is: jerk
negetive_word_is: shame
total_negative::4
Result:: positive sentiments
```

Fig. 4 Analysis output

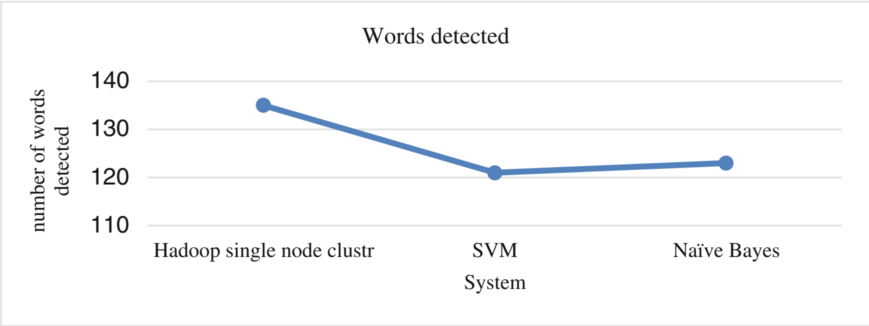


Fig. 5 Comparison of detected words

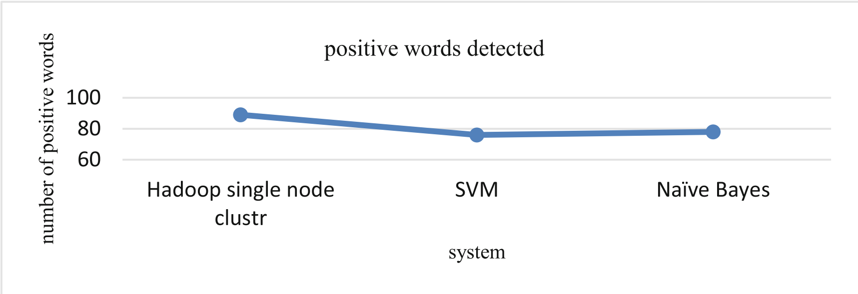


Fig. 6 Positive words detected

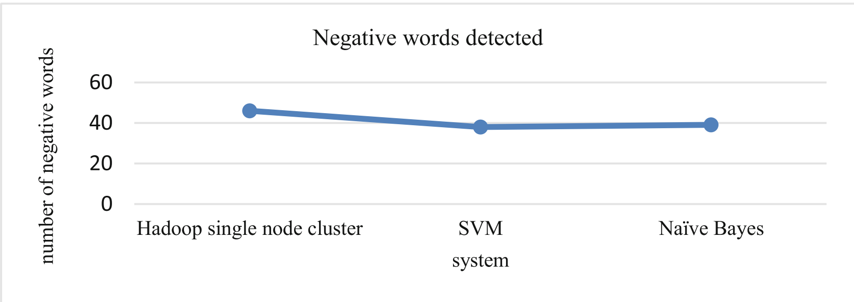


Fig. 7 Comparison of detected negative word

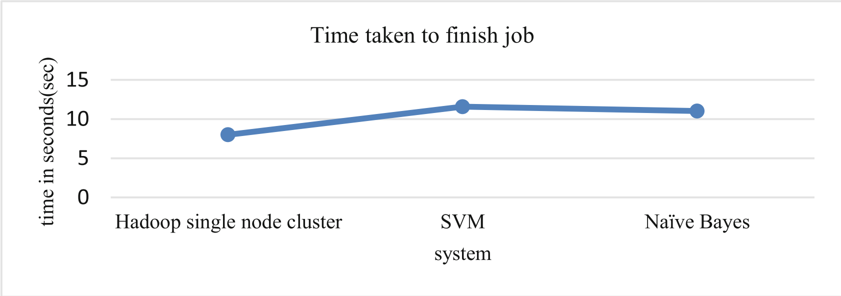


Fig. 8 Comparison of time taken to finish the job

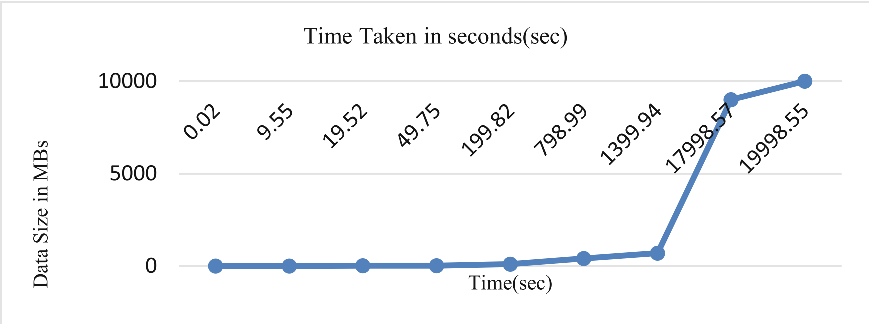


Fig. 9 Time graph

Table 2 Result analysis table

S. no.	System	No. of detected words	No. of positive words	No. of negative words	Time is taken to finish job (s)
1	Hadoop single node cluster	135	89	46	7.99
2	SVM	121	76	38	11.58
3	Naïve Bayes	123	78	39	11.02

7 Conclusion

The focus of the paper was on improving the time consumed in analyzing the sentiments of the people. The results clearly indicate that the goal was successfully achieved. The proposed system consumed much lesser time when compared with

Table 3 Time taken to perform jobs on different data size

S. no	Data size (MBs)	Time taken to finish job (s)
1	1	0.02
2	500	9.55
3	1000	19.52
4	2500	49.75
5	10000	199.82
6	40000	798.99
7	70000	1399.94
8	90000	17998.57
9	1000000	19998.55

SVM and Naïve Bayes. The proposed system consumed 45% less time over SVM and 38% less than Naïve Bayes. The accuracy in terms of a total number of words detected, positive and negative words was 11%, 16%, 18% respectively better than SVM and 9%, 13%, 16% respectively better than Naïve Bayes. It can be concluded the proposed system performed much better than SVM and Naïve Bayes. The authors are working on implementing the system on heterogeneous Hadoop clusters over GPU in future to better the results observed in the proposed work.

References

1. Lam, C., Davis, M., Gaddam, A.: Hadoop in Action, 2nd edn. Manning Publications (2016)
2. Kenekar, T.V., Dani, A.R.: An efficient private FIM on hadoop MapReduce. In: IEEE International Conference on Automatic Control and Dynamic Optimization Techniques, Pune, India, Sept 9–10, 2017. https://doi.org/10.1109/icacdot.2016.7877554
3. Kumar, S., Singh, P., Rani, S.: Sentimental analysis of social media using R language and Hadoop: Rhadoop. In: 5th IEEE International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), Noida, India, Sept 7–9 (2016). https://doi.org/10.1109/icrito.2016.7784953
4. Islam, M.R., Zibran, M.F.: A comparison of dictionary building methods for sentiment analysis in software engineering text. In: Empirical Software Engineering and Measurement, Toronto, ON, Canada, Nov 9–10 (2017). https://doi.org/10.1109/esem.2017.67
5. Ameur, H., Jamoussi, S.: Dynamic construction of dictionaries for sentiment classification. In: 13th IEEE International Conference on Data Mining Workshop, Dallas, TX, USA, Dec 7–10 (2013). https://doi.org/10.1109/icdmw.2013.34
6. Mallika, C., Selvamuthukumaran, S.: Hadoop framework: analyzes workload prediction of data from cloud computing. In: IEEE International Conference on IoT and Application, Nagapattinam, India, May 19–20 (2017). https://doi.org/10.1109/iciota.2017.8073624
7. Ahmed, K., Tazi, N.E., Hossny, A.H.: Sentiment analysis over social networks: an overview. In: IEEE International Conference on Systems, Man, and Cybernetics, Kowloon, China, Oct 9–12 (2015). https://doi.org/10.1109/smc.2015.380
8. Nanli, Z., Ping, Z., Weiguo, L.: Sentiment analysis: a literature review. In: IEEE International Symposium on Management of Technology, Hangzhou, China, Nov 8–9 (2012). https://doi.org/10.1109/ismot.2012.6679538

9. Povoda, L., Burget, R., Dutta, M.K.: Sentiment analysis based on support vector machine and big data. In: IEEE 39th International Conference on Telecommunications and Signal Processing, Vienna, Austria, June 27–29 (2016). https://doi.org/10.1109/tsp.2016.7760939
10. Ahlgren, O.: Research on sentiment analysis: the first decade. In: 16th IEEE International Conference on Data Mining Workshops, Barcelona, Spain, Dec 12–15 (2016). https://doi.org/10.1109/icdmw.2016.0131
11. Chandankhede, C., Devle, P., Waskar, A.: ISAR: Implicit sentiment analysis of user reviews. In: International Conference on Computing, Analytics and Security Trends, Puna, India, Dec 19–21 (2016). https://doi.org/10.1109/cast.2016.7914994

Multi-GPU Approach for Development of Parallel and Scalable Pub-Sub System

Medha A. Shah and Dinesh Kulkarni

Abstract Event matching plays an important part in the overall attainment of the content-based Publish-Subscribe system. These systems demand guaranteed message delivery, high throughput and low matching time. Existing parallel content matching algorithms make use of multiple cores and off the shelf hardware easily available in today's modern computers. For a large number of events and subscriptions, these algorithms suffer from performance degradation. In this paper, we propose high-performance Publish-Subscribe system designed to run efficiently on multiple GPUs. Performance comparison with existing CCM (CUDA Content Matching) algorithm clearly demonstrates 32% improvement in matching latency.

Keywords High performance · Parallel matching algorithm · Multiple GPUs
Matching latency

1 Introduction

Many distributed applications use Publish-Subscribe (Pub-Sub) communication paradigm as communication backbone. In the Pub-Sub model, subscribers typically receive only a subset of the messages published by one or more publishers. Here receivers declare their interest in the particular event in the form of subscription. The publisher publishes the information of interest as messages or notifications. The content-based Pub-Sub system delivers to the subscribers published messages, which matches subscriber's declared interest. The efficient matching of an event with a large number of subscribers on a broker of a distributed system is a fundamental issue to be

M. A. Shah (✉)
Department of CSE, Walchand College of Engineering, Sangli, Maharashtra, India
e-mail: medha.shah@walchandsangli.ac.in; shah.medha@gmail.com

D. Kulkarni
Department of Information Technology, Walchand College of Engineering, Sangli, Maharashtra, India
e-mail: dinesh.kulkarni@walchandsangli.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_49

focused. To minimize matching latency and to deliver high throughput (events processed per second) are two fundamental goals of the Pub-Sub system. Content-based applications like intrusion detection system, high-frequency trading requires very low response time. Earlier developed matching algorithms are inherently sequential and hence unable to take advantage of current parallel architecture easily available in modern computers. It is necessary to design and develop scalable parallel matching algorithms which can exploit Multi-core, Many-core and accelerator frameworks to obtain higher matching performance.

There are recent studies on the development of a high-performance content-based system using GPU [1]. As an example, CCM algorithm, proposed in [1] achieved low matching latency and impressive speedup of 66.7 while compared with SFF. This work [1] commonly assumed that all subscriptions can fit into the memory of single GPU. The size of host memory is much larger than the device memory. When dealing with data-intensive applications, the size of the device memory may thus become a limiting factor. Here our aim is to improve the performance of matching algorithm for data-intensive applications, by overcoming device memory limitation using multiple GPUs.

Moving to our research contributions, we developed two Multi-GPU Cuda Content Matching algorithms namely MCCM (Single Event) and MCCM (Parallel Events). The work presented earlier in paper [2] talks about design of MCCM (Parallel Events) algorithm. Both algorithms leverage multiple GPUs to reduce overall matching latency. We study the algorithm performance in comparison with CCM, a high performance, parallel content matching algorithm. This study demonstrates that leveraging multiple GPUs can bring impressive improvement in matching latency. This paper is organized as follows. Section 2 presents related work in the area of High-Performance Pub-Sub system. Section 3 presents data model typically assumed for the content-based Pub-Sub system. Section 4 describes parallel matching algorithm and its implementation on Multiple GPUs. Section 5 shows experimental setup and comparative results of all approaches.

2 Literature Review

In [3], the author proposed the idea of parallel matching by exploiting multi-core CPUs. In this paper, two techniques namely single event collaborative processing and multiple events independent processing are proposed. The first technique aims at reducing the processing time of a single event and second technique aims at reducing overall matching latency. Due to limited availability of threads above mentioned approaches cannot exhibit higher performance. The use of ad hoc (FPGA) hardware is shown in the parallelization of matching process in [4]. The Margara in [1] proposed high-performance content-based matching algorithm using GPUs and multi-core CPUs. The algorithm requires a substantial amount of time for processing large numbers of subscriptions, filters and interfaces. Stream Hub [5] is a novel Pub-Sub deployed on a cluster of 384 cores. This Pub-Sub is able to register and filter a large number of publications against stored subscriptions resulting nearly 400 K notification/s. The quality of Service in Pub-Sub system is maintained by designing dead line aware algorithm mentioned in [6]. Paper [2] presents and validates design of high-performance Pub-Sub system by using parallel platforms like OpenMP and CUDA and also by using some real time event processing frameworks like Storm. MPI-CUDA approach for design of high-performance system is discussed in [7].

3 Data Model

Here we illustrate what we mean by event, filter and predicate. An event is a set of attributes. Each attribute has a name, type and a value. For example, string class = travel/airlines/offer; date starts = Jun; date expires = Aug; string origin = LA; string destination = AUS; string carrier = United is an event. An event is also defined as an attribute value pair. A Filter is defined as the conjunction of attribute constraints. Each attribute constraint has a name, a type, an operator, and a value. A constraint defines an elementary condition over an event or message.

4 Multi-GPU Parallel Content Matching Algorithms

This section presents our MCCM (Single Event) approach and its implementation on Multi-GPU system. The key goal in designing MCCM (Single Event) is to process the single event collaboratively by multiple GPUs to reduce matching latency of single event. MCCM (Parallel Events) algorithm [2] focuses on increasing the throughput. In this approach multiple events are processed simultaneously which reduces overall matching latency and thus increases the throughput. Here each GPU processes individual event and returns result back to the CPU. This approach requires minimal synchronization. Proposed algorithms make use of CCM algorithm presented in [1] as a base algorithm. Here we briefly explain CCM algorithm.

4.1 CCM Algorithm

CCM is parallel content matching algorithm designed to run efficiently on single GPU. This is based on counting algorithm [8, 9], prominently used for matching in Pub-Sub system. The algorithm works in three phases. The first phase is a filter selection phase which is followed by a constraint selection phase. Next step is constraint evaluation phase and finally counting phase is accomplished [1]. In the filter selection phase, the set of filters is divided based on their attributes names. In the second phase, for each individual attribute a, in the event e, the set of constraints (part of the filters) having the same name as a is selected. Evaluation of each selected constraint is carried out in the last phase using the value of each attribute a. For each satisfied constraint “c”, a counter associated with filter f is increased. When all the constraints of the filter are satisfied, a filter matches an event e and so satisfies the predicate p to which it belongs. When this process is completed, an event can be directed to the interface subject to predicate “p”. CCM algorithm evaluates multiple constraints in the filters simultaneously, by using GPU cores. The algorithm maintains following data structures for storage of events and subscriptions. Here onwards CPU is referred as host process. Host process maps each distinct attribute name to a single bit of small bit vector called NameVector (NVe). It also creates Filters and Constraints tables, to organize the filter constraints into five main data structures namely ConstrOp, ConstrVal, ConstrFilterId, ConstrBF, and Num-Constr [1], according to constraint name. Host process builds data structures named Filters and Interface [1] to maintain information about filter-size, filter-count, and interface-Id to which the filter belongs. An Interface is an array, indicates matched interfaces by setting the corresponding index of the array to one. For each event, host process builds the table Input. It includes one row for each attribute in event e. For each attribute a, in event e, each row of the input table stores the value of a, its type, the number of constraints having the same name as a, and the pointers (in the GPU memory) to the rows of Filters and Constraint table that are relevant for a. All these data structures are transferred to the GPU for processing an event. CCM launches a kernel named evalConstraint, which exploits thousands of GPU threads for parallel evaluation of multiple constraints. A single attribute a of event e is evaluated against a single constraint c by each thread. The results of the computation (i.e., the Interfaces array) are copied back to the host memory and the FiltersCount and Interfaces structures are reset for processing of the next event.

4.2 MCCM (Single Event)

The aim of this algorithm is to match a single event collaboratively by multiple GPUs to improve matching latency of a single event. The overall idea of this approach is like this. Total interfaces through which subscriptions (predicates) are entered in the Pub-Sub system are partitioned equally and assigned to available GPUs for processing. Here each GPU will perform the task of matching the same event with predicates assigned to it. Partial results are generated by each GPU. Finally, partial results are combined to produce the output of the algorithm. This approach helps to solve limitation on data size posed by the memory of single GPU. This is possible because approximately half of the total interfaces are assigned to individual GPU for processing. MCCM builds same data structures as in CCM but in addition to this, algorithm manipulates interface array on host differently. Each GPU will simultaneously update matched interfaces for given event in an interface array. Here onwards term device refers to GPU in compute node. Throughout the algorithm, following terminologies are used. Device input and device interface refer to respective input and interface array on the device. Host interface refers to interface array on the host.

The term offset refers to a location in interface array on the host, for copying matched interfaces. This algorithm uses same constraint evaluation kernel presented in [1]. The pseudocode of the MCCM (Single Event) is presented in Algorithm 1. At the first two lines, Interfaces are divided and assigned to GPUs and Constraints and Filters tables on the host are transferred to the memory of every GPU asynchronously using default stream of GPU. At line 3 and 4 incoming events are kept in Input stream and Input table for the event has been created. At line 5, single CPU thread manages the multiple GPUs. The cudaSetDevice function sets the current GPU and CUDA calls are issued to the current device. Here each device issues calls to constraint evaluation kernel asynchronously. At line 6, the partial results generated by each GPU are copied to interface array maintained at host process at the correct location. Finally, synchronization of devices takes place.

Algorithm 1 MCCM (Single Event)

```
1. Partition and assign Interfaces (channels, exposed to predicates) on host process to GPUs
   available for processing.
2. Create Filters and Constraint tables for Interfaces on host process and transfer to respective
   GPGPU memory.
3. Store incoming events to Input stream.
4. Create Input table (host_input) for an event to be processed.
5. for each GPGPU 'i' do
          cudaSetDevice(i)
          cudaMemcpyAsync(device_input, host_input, cudaMemcpyHostToDevice)
          evalConstraint<<<NUM_BLOCKS, NUM_THREADS>>>
   end for
6. for each GPGPU 'i' do
          cudaSetDevice(i)
          cudaMemcpyAsync(host_interface+i*offset,device_interface, cudaMemcpyDeviceTo-
   Host)
          cudaDeviceSynchronize()
          cudaMemset(device_interface)
   end for
```

5 Evaluation

The performance of the algorithm is evaluated using synthetic workloads generated from auxiliary workload generator mentioned in [1]. The intention behind conducting various experiments was to prove benefits of two-level parallelism for designing high-performance content-based Pub-Sub system.

5.1 Experimental Setup and Parameters

For evaluating Multi-GPU parallel content matching algorithms, Multi-GPU framework has been setup. Multi-GPU system is homogeneous. Two GPU cards are of

Table 1 Processing time in the default scenario

CCM	MCCM (Single Event)	MCCM (Parallel Events)	SFF
0.25245 ms	0.1798 ms	0.1505 ms	1.035 ms

same specifications. All tests of this setup were executed on a Core i7-2600 PC, with four cores running at 3.2 GHz, and 16 GB of DDR3 Ram. The Two GPGPUs were a NVIDIA Quadro K600 with 1 GB of DDR3 RAM. It has 192 SMX CUDA parallel processing cores. We used the GCC compiler, version 4.7, and the CUDA runtime 7.5 for 64 bit Linux. Both the GPUs are residing on a single node. To analyze and optimize the performance of our algorithm we have used NVIDIA profiling tool nvprof. To test the algorithm in this framework, which uses two GPUs following command is used.

$$\text{Nvprof}-\text{print} - \text{gpu} - \text{trace}$$

This command shows on which GPU each kernel is running as well as grid dimensions used for each kernel launch. Another visual profiler used during experimentation is nvvp which displays activity of your application on both CPU and GPU. Some performance improvement was made looking to the results obtained from nvvp.

5.2 Matching Latency

The processing time required by all the algorithms is shown in Table 1. Algorithms use workload mentioned in default scenario of [1]. Under this load, CCM requires 0.25245 ms to process a single event, our Multi-GPU algorithm namely MCCM (Single Event) requires 0.1798 ms and MCCM (Parallel Events) requires 0.1505 ms providing respectively an improvement of 40% and 32% on CCM. SFF (Siena Fast Forwarding algorithm) requires 1.035 ms for matching. This algorithm requires long time due to its sequential nature.

Here MCCM (Single Event) requires more processing time due to collaboration required by both the GPUs for processing a single event. Here synchronization is required at some point. But eventually, MCCM (Single Event) approach works well for a large number of Attributes.

Numbers of Interfaces

Figure 1 shows the performance of algorithms with increasing number of interfaces. It is observed that CCM is unable to process a large number of interfaces. The reason may be memory limit of single GPU. MCCM (Single Event) algorithm, collaborates to process the single event by distributing interfaces among available Multiple GPUs on node. As Interfaces are divided among two GPUs, the algorithm is able to produce

Fig. 1 Numbers of interfaces

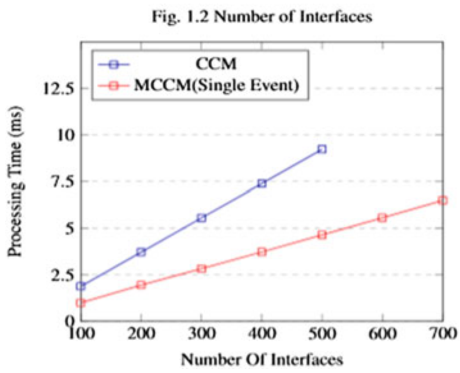
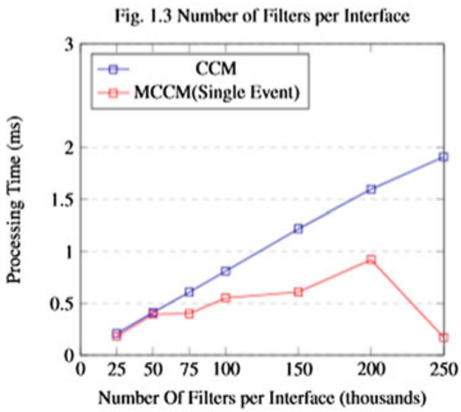


Fig. 2 Number of filters per interface



the results for a large number of interfaces. Nearly 1.5X improvement in matching latency in comparison with single GPU CCM algorithm is observed.

Figure 2 shows impact of changing number of filters per interface. With increase in filter-count, overall number of constraints also get increases leading towards more complex scenario. This scenario emphasizes the advantages of collaborative event processing in MCCM (Single Event) approach. In this approach as single event is collaboratively processed by multiple GPUs, interfaces are divided among available GPUs. In turn filters get divided and MCCM (Single Event) approach exhibits better performance in comparison with other approach.

5.3 *Final Considerations*

Parallel, scalable and high-performance Pub-Sub system can be designed by extending single GPU based CCM algorithm. It is possible to speedup data-intensive application using multi-GPU system. Multi-GPU feature becomes significant when work-

Table 2 Comparison of pub-sub systems

Parameter	OpenMP	Multi-GPU	MPI-CUDA	Storm	SFF
Matching Latency (ms)		Minimum (0.1505)			Maximum (1.035)
Communication overhead			Maximum		
Throughput				Maximum (7000)	
Speedup		Maximum (112 X)			
Data-intensive applications		Recommended	Recommended		
Scalability	limited		Maximum	Maximum	

load or input exceeds the capacity of single GPU. CUDA makes it easy to manage multiple GPUs using a single host thread MCCM (Single Event) algorithm requires more synchronization while processing. This approach is useful for a large number of subscriptions. Table 2 compares matching approaches used in Pub-Sub systems on various parameters.

References

1. Margara, A., Cugola, G.: High performance publish-subscribe matching using parallel hardware. In: IEEE Trans. Parallel Distrib. Syst. **25** (2014)
2. Shah, M.A., Kulkarni, D.B.: Doctoral symposium: design and development of high performance, scalable content based publish subscribe system. In: Proceedings of the 10th ACM International Conference on Distributed and Event-based Systems (2016)
3. Farroukh, A., Ferzli, E., Tajuddin, N., Jacobsen, H.-A.: Parallel event processing for content-based publish/subscribe systems. In: DEBS vol. 09, pp. 8:1–8:4 (2009)
4. Tsoi, K., Papagiannis, I., Migliavacca, M., Luk, W., Pietzuch, P.: Accelerating publish/subscribe matching on reconfigurable supercomputing platforms. In: Proceedings of Many-Core and Reconfigurable Supercomputing Conference
5. Barazzutti, R., Felber, P.: Streamhub: A massively parallel architecture for high-performance content-based publish/subscribe. In: DEBS vol. 13
6. Wang, X.C.: Pub/sub on stream: a multi-core based message broker with qos support. In: DEBS (2012)
7. Shah, M.A., Kulkarni, D.B.: Design of high performance, scalable content-based publish-subscribe system using MPI-CUDA approach. In: Int. J. Comput. Sci. Eng. (IJCSE) UGC approved (2017). ISSN: 2347–2693 (E)
8. Carzaniga, A., Wolf, A.: Forwarding in a content-based network. In: Proceedings of SIG-COMM, pp. 163–174 (2003)
9. Carzaniga, A., Rosenblum, D., Wolf, A.: Achieving scalability and expressiveness in an internet-scale event notification service. In: Proceedings of the 19th Annual ACM Symposium, pp. 219–227 (2000)

Artificially Talented Architecture for Theme Detection

A. Karamchandani, T. Agey, A. Chavan, Vaibhav Khatavkar and Parag Kulkarni

Abstract Intelligent systems are the need of today's world. Collections of data and various data sets are made available to naive users. Understanding what is contained within the dataset is quite difficult by referring just the name. Some of the datasets have quite a difficult, weird names so users do not have any clue what is inside, so there is a need of the theme of the document or dataset so as to understand what are the contents. User satisfaction and convenience is of prime importance. In this paper, we try to propose a system along with a working prototype of such intelligent system that essentially is a Chatbot which uses facility of Theme Detection in semantic analysis stage while processing the user input. This makes the system more productive. This paper talks about Chatbot and improvement in intelligent responses using theme detection. We have built a prototype of the system.

Keywords Intelligent system · Chatbot · Theme detection · Semantic analysis
Context vector · Text analysis · Vector space model

A. Karamchandani (✉) · T. Agey · A. Chavan · V. Khatavkar · P. Kulkarni
Department of Computer Engineering and IT, College of Engineering, Pune, MH, India
e-mail: adityabk1996@gmail.com

T. Agey
e-mail: agey.tushar3@gmail.com

A. Chavan
e-mail: cajinkya21@gmail.com

V. Khatavkar
e-mail: vkk.comp@coep.ac.in

P. Kulkarni
e-mail: paragindia@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_50

1 Introduction

In current markets, we see many commercial implementations of chatbots. Creating a rule-based chatbot is easy, but when it comes to the development of intelligent conversation systems, more advancement is required. Natural language used by human user in a chatbot scenario can be ambiguous and reducing this ambiguity is a tricky job.

The idea behind the theme detection using an intelligent system is basically the integration of user interface with data mining capabilities. Due to vagueness of language, many a times it so happens that a word can have different meaning based on context and to identify the meaning in certain context, we detect the theme [1, 2]. Along with Theme Detection, we have idealized a chatbot after looking at the user needs of a good and friendly interface for chatbot. Artificially Talented Architecture (ATA) is used for taking speech input and produce speech output. This allows user to speak what to do and the system will act accordingly. Hence, its a voice [3] command system.

To solve this problem, we have build a chatbot application that has unique holographic [4–6] user interface. The motivation behind the proposed work is the paper on Context Vector Machine [7]. In their paper, they have mentioned a way of building a parse tree for a set of documents and process it so as to find context. They also have mentioned that it is possible to use audio/voice as an input. Hence we have used a chatbot [ATA] for taking audio input and identify theme of the document.

The chatbot takes the voice input from user and processes the input at server side. This input is then matched against the conversational dataset using similarity measure and the output is given back to the user as speech. The theme identification module also gets the input and according to theme, the output is modeled to best suite the situation.

Section 2 describes the literature review followed by Document Analysis in Sect. 3. Section 4 describes the use of "Theme analysis" in Chatbot communication with subsections which elaborates on intermediate modules. In Sect. 5, we conclude with the benefits and future work that can be done in it.

2 Literature Review

Chatbots are generally built or implemented on IRC channels, to maintain and manage them.

Rodrigo et al. [8] state other uses of chatbots on Social Media platforms as

1. Advertising (Spam)
2. Entertainment
3. Customer Service (Knowledge Database).

Another type of Artificially Intelligent chatbots are the standalone chatbots, which provide functionality other than the Social Media chatbots, some are listed below.

1. Planning and Organization
2. Email and Call Management
3. Conversation with the user
4. Fetch data from Internet.

Artificial Linguistic Internet Computer Entity (ALICE) [9] is the first Artificial Intelligence Markup Language (AIML) based personality program that pretends to be intelligent and self aware. On the other hand Siri [10], Google Assistant [11] come with their respective devices, are actually better when it comes to the understanding the Natural Language and interacting with the user. Amazon Alexa is Amazon's approach to build AI chatbot, which can also help the user with ordering products from Amazon.com [12]. The average IQ of these chatbot assistants is also pretty impressive.

When it comes to Artificially Talented Architecture (ATA), it is an interface with a inverted pyramid on top of it, so as to get a Hologram of the bot's user interface. Here the outer surfaces of the inverted pyramidal prism, refract the light and when all the 4 different images converge in the middle of the inverted pyramid then it looks as a floating hologram. Also since Duck Duck Go API [13] is used, there is no problem of User's being concerned with their Personal Information being tracked, as there is no such tracking done by Duck Duck Go. Neither is there a concept of filter bubble [14], so that irrespective who the user is, they would get proper and uniform answers and not the one that is tweaked and changes from person to person.

Also ATA could interact with humans and tell them jokes and take appointments and make note of what user specifically asks ATA to remember. ATA is also scalable and the further extension of ATA could be as an Interface to the Theme Detection Agent of a Document. ATA would take the input Document/(s) and come up with the theme as the output.

So, let us have a look at the Vector Space Model (VSM) and the Inner Product Similarity Measure.

Similarity measure: Let us say the query vector is,

$$\vec{q} = (q1 * w1 + q2 * w2 + q3 * w3 + \ldots),$$

where, $q_1, q_2, q_{3,}$ .... are the individual independent terms which constitute the Basis Vectors of the VSM and $w_1, w_2, w_{3,}$ ... are their respective weights. These weights can either be binary, saying whether the term is present in the query vector or not, or can be real valued, which can represent how many times does the term appear in the vector.

Now, similarly all the documents on the server are already stored as their Vector forms in the Vector Space Model (VSM). So, the documents would look like

$$\overrightarrow{doc1} = (t1 * w1 + t2 * w2 + \ldots)$$
$$\overrightarrow{doc2} = (t1 * w1 + t2 * w2 + \ldots)$$
$$\overrightarrow{doc3} = (t1 * w1 + t2 * w2 + \ldots)$$
$$\vdots$$
$$\overrightarrow{docn} = (t1 * w1 + t2 * w2 + \ldots)$$

So now, to calculate the relevancy between two vectors $\overrightarrow{q}$, $\overrightarrow{doc_k}$ we use Inner Product Similarity Measure. Inner Product Similarity is calculated using the following formula between two vectors

$\vec{x}$, $\vec{y}$, for z being the length of the vectors

$$sim\left(\overrightarrow{x}, \overrightarrow{y}\right) = \sum_{n=1}^{z} x_i * y_i$$

for calculating the similarity between the query and some document k, we'll use

$$sim\left(\overrightarrow{q}, \overrightarrow{doc_k}\right) = \sum_{n=1}^{z} q_i * doc_{k,i}$$

So now, doing this calculation for all the documents in the corpus we can find out the most similar documents to the query in the corpus.

Takuwa et al. [15] have emphasized upon the importance of Document Classification and have proposed a new method for the same using Bag of Words (BOW) and multiple distributed representations. Qian et al. [16] have given a scientific theme emergence detection approach based on Citation Graph Analysis, as currently the mainly performed operation on the contents of the files are Topic Detection. Therefore they have used the topic emergence detection through the comparison structure of the structure of the community over time.

Communities are modeled in the citation graph as related documents on a specific topic, then scientific theme detection algorithm is proposed based on the community partition, and an attempt is made to identify the emergence of the new theme by tracking the change of the community where the top cited nodes lie in. Experimental results have been shown to detect the new topic timely even with only a subset of the data.

Bawakid [1] have proposed a new unsupervised approach for identifying the main themes of any text document with the aid of Wikipedia. In contrast to others, the proposed algorithm relies on merely two main aspects of Wikipedia, namely its articles' titles and categories structure. There is a slight limitation to this, as the content of the articles are not employed anywhere in the process.

Liu et al. [17] have concentrated on the comment panels of various Massive open online course (MOOC). MOOCs have played a very important role in the lives

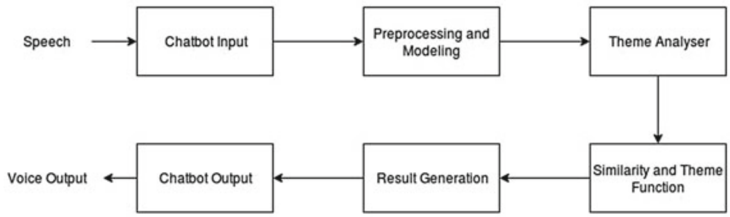


Fig. 1 Architecture diagram

of students and teachers. MOOCs offer a variety of interactive ways, in which the comment panel is used for expressing students' opinions and feelings. It is imperative that these feedbacks need to be registered and should be analyzed for the detection of context of the commenter. Li et al. [18] deal with topic detection for forums. Forums have become one of the main platforms for people to express their point of view. So how to detect a forum topic automatically based on the forum contents and the massive information has become an important task. Agglomerative Hierarchical Clustering (AHC) is used for solving this problem. Given the topic a slight idea of the theme of the document can also be achieved.

Nassar et al. [19] used the Twitter Exemplar-based topic detection system and to improve its quality. The feedback from the crowd is utilized to adjust weights of the cosine similarity function deployed in the Exemplar-based topic detection algorithm. Since it is crowd sourced there is a very great need of managing the project. Also volunteered participation is expected. Results are based on the dataset used and generalized. Chen et al. [20] solved a major issue in Natural Language Processing and concentrates on the extracting of the Event or the Topic of the corpus. Limitations are, the connections among features are lost in transformation and representation from text to vectors, because of the limitation of independence assumption of Vector Space Model (VSM).

3 Proposed System

The chatbot takes user input as a speech and converts it into the text. This text is then given to Theme analysis module which identifies the theme of the text. Then detected theme and the original text is sent to the server for further analysis. Server refines its dataset [21] based on the Theme and then finds out the most similar result. This result is then sent back. This result is then converted to speech and user can perceive the result (Fig. 1).

3.1 Speech to Text Input

This module takes input from user in the form of speech. Then with the help of speech recognition technique, system converts the audio into the text. Theme analyzer. This module uses Context Vector Machine [7] for identifying the theme of the sentence spoken by the user. This gives insights to the chatbot of what the user is talking about, which makes the chatbot smarter.

3.2 Similarity and Theme Function

This module is a centralized module residing over a server that we have set up. This server module is written in PHP and javascript. This module takes the theme and the input text and tries to find the best matching answer based on the theme and the input text by using similarity function of inner product with the knowledge base that it has.

3.3 Result Generation and Verification

This module generated the final result and if its satisfactory, then forwards it to the chatbot client.

3.4 Chatbot Output

This module takes the data coming from server and speaks it out.
Algorithm

Step1: text = STT (user_voice_input)
Step2: theme = identify_theme (text)
Step3: result = server_process (text, theme)
Step4: TTS (result)

Algorithm 1: Overall System Design.

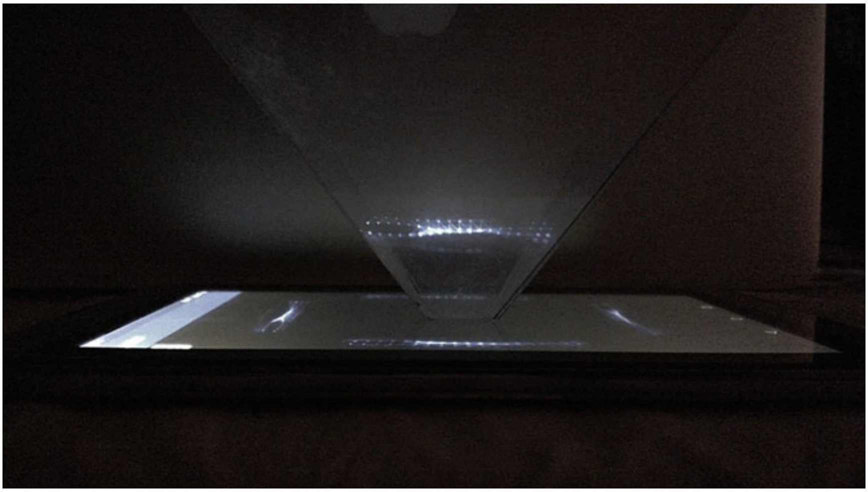


Fig. 2 Working module with holographic user interface

4 Results and Discussion

We have ATA (Artificially Talented Architecture) built as a mobile application. This mobile application has a view containing four interdependent panels. These panels run User Interface at an instance of time. At the center of the screen, a Pyramid is inverted and placed. This gives fabulous effect of Hologram [22]. All the images on four sides of the pyramid converge using refraction and gives the feeling as if the User Interface (UI) is floating in the air. This makes the system unique in visuals as well as unique in behavior (Fig. 2).

4.1 Theme and Similarity Analysis

The dataset used is communication dataset from chatterbot-corpus [23]. This dataset is used for communication and theme analysis of a dialogue making the conversation more context-oriented. The Artificially Talented Architecture (ATA) when takes the input from the user, it sends the input and the theme of the document to the central server for further processing. The theme analysis module first preprocesses the text by removing stop words and punctuations. This makes the query lightweight and easy to process. We use *nltk.parse.stanford* [24] package from the Python Library and import the module *StanfordDependencyParser* [25] to generate a dependency parse tree. Now this Parse tree is used to create a Matrix, upon which we will perform operations and figure out the Theme of the Document. This Theme and the original query is then sent to the Server. Server represents the query as a vector in the Vector

Table 1 Comparative analysis of ATA for theme detection with existing methods

Sr. no.	Paper	Idea	Comment
1.	Wikipedia categories for discovering the themes of text documents [1]	Identification of theme using title of article and Wikipedia tags	Contents of documents is not taken into consideration
2.	Context vector machines for information retrieval [7]	Proposed system would classify a document using multiple classification engines	The system proposed in this paper is an extension of this work
3.	Theme detection an exploration of opinion subjectivity [2]	Theme detection technique classifies the sentences into one of the two binary features: opinionated and nonopinionated. Sentence level opinions are accumulated and subjectivity is obtained	Precision and recall of the overall system in the proposed system is 76.08% and 83.33% respectively for the English language. Can be improved in the future versions of this work
4.	Google assistant [11]	It is the personal assistant developed by Google Co. which acts as a personal helper to users via a mobile interface	There is no attractive or holographic UI as developed in the proposed system

Space Model (VSM) of all the unique terms in the index of the document corpus inverted index stored already on the server. The search space on the server is confined by the theme of the document. Then most similar result is sent to the user which as a voice output (Table 1).

5 Conclusion

In Artificially Talented Architecture for Theme Detection (ATA) the introduction starts with the discussion of chatbots and currently present functionalities. How ATA adds on to the functionalities is explained in Sect. 3. Then there is literature review which explains the working of existing chatbots [26, 27]. Then we have Proposed System with architecture and overall algorithm of the system. Mathematical Model talks about similarity measure and theme analysis. So a theme detection is done along with holographic user interface which accepts the input as speech and returns the results as speech output so it is an intelligent system which interact with the user as a human and also detects the theme of the document mentioned as input.

In future, a system can be developed which accepts the whole dataset and do the analysis of theme as well as we can enhance the user interface as per the need of the era. The similarity measure can be improved using higher and complex mathematical models and data mining concepts for better results.

References

1. Bawakid, A.: Using wikipedia categories for discovering the themes of text documents. In: 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, vol. 1, pp. 452–455, Aug 2015
2. Das, A., Bandyopadhyay, S.: Theme detection an exploration of opinion subjectivity. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–6, Sept 2009
3. Quintero, J., Asprilla, R.: Towards an efficient voice-based chatbot. In: 2015 IEEE Institute of Electrical and Electronics Engineers Thirty Fifth Central American and Panama Convention (CONCAPAN XXXV), pp. 1–6, Nov 2015
4. Angga, P.A., Fachri, W.E., Elevanita, A., Suryadi, Agushinta, R.D.: Design of chatbot with 3D avatar, voice interface, and facial expression. In: 2015 International Conference on Science in Information Technology (ICSITech), pp. 326–330, Oct 2015
5. Kaushik, D., Kumar, D: A review paper on holographic projection. IJIRT Int. J. Innov. Res. Technol. **1**(6), 1–8 (2014)
6. Richardson, M.J., Wiltshire, J.D.: What is a Hologram? Wiley, pp. 336. IEEE—Institute of Electrical and Electronics Engineers—Press, (2018)
7. Kulkarni, P., Khatavkar, V.: Context Vector Machine for Information Retrieval, vol. 137
8. Rodrigo, S.M., Abraham, J.G.F.: Development and implementation of a chat bot in a social network. pp. 751–755, Apr 2012
9. Wallace, R.S.: The Anatomy of A.L.I.C.E., pp. 181–210. Springer Netherlands, Dordrecht (2009)
10. Apple Inc. iOS—Siri—Apple (2017). https://www.apple.com/ios/siri/. Accessed 05 Dec 2017
11. Google. Google assistant—your own personal Google (2017). https://assistant.google.com/intl/en_in/. Accessed 05 Dec 2017
12. Amazon (2017). https://amazon.com/. Accessed 05 Dec 2017
13. DuckDuckGo Inc. (2017). https://duckduckgo.com/. Accessed 05 Dec 2017
14. Pariser, E.: The Filter Bubble: What the Internet is Hiding from You, vol. 137
15. Takuwa, K., Yoshikawa, T., Jimenez, F., Furuhashi, T.: A study on document classification using multiple distributed representations. pp. 1–4, June 2017
16. Qian, T., Sheu, P.C.Y., Li, S., Wang, L.: A scientific theme emergence detection approach based on citation graph analysis. vol. 2, pp. 269–273, Nov 2008
17. Liu, Z., Zhang, W., Sun, J., Cheng, H.N.H., Peng, X., Liu, S.: Emotion and associated topic detection for course comments in a MOOC platform. pp. 15–19, Sept 2016
18. Li, H., Li, Q.: Forum topic detection based on hierarchical clustering. pp. 529–533, July 2016
19. Nassar, L., Ibrahim, R., Karray, F.: Enhancing topic detection in twitter using the crowdsourcing process. pp. 196–203, Oct 2016
20. Chen, Y., Liu, L.: Development and research of topic detection and tracking, pp. 170–173, Aug 2016
21. Wibowo, F.W., Setiaji, B.: Chatbot using a knowledge in database. In: 7th International Conference on Intelligent Systems, Modelling and Simulation, vol. 2016
22. Manekiya, M.H., Arulmozhivarman, P.: 3D volume reconstruction using hologram. In: 2016 International Conference on Communication and Signal Processing (ICCSP), pp. 1570– 1574, Apr 2016
23. gunthercox. GitHub—gunthercox/chatterbot-corpus: a multilingual dialog corpus (2017). https://github.com/gunthercox/chatterbot-corpus. Accessed 20 Dec 2017
24. Loper, E., Bird, S., Klein, E.: Natural language toolkit—NLTK 3.2.5 documentation (2007). http://www.nltk.org/. Accessed 20 Dec 2017

25. Stanford NLP Group. The Stanford Natural Language Processing Group. https://nlp.stanford.edu/. Accessed 20 Dec 2017
26. Behera, B.: Chappie—a semi-automatic intelligent chatbot
27. Setiaji, B., Wibowo, F.W.: Chatbot using a knowledge in database: human-to-machine conversation modeling. In: 2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS), pp. 72–77, Jan 2016

Study and Effect of Architecture Deployed in BPO on Screen Recording Compliance for In-Centre Versus at-Home Agents

Rajendra Deshpande, Ulhas Shiurkar and Satish Devane

Abstract In business process outsourcing (BPO), Voice and Screen recording systems are deployed. Deploying these systems addresses the varied BPO needs such as compliance, quality management, improving customer experience, and dispute resolution. As per the architecture, recording systems record the agent and the customer voice. In addition, they screen capture the transactions carried out by an associate on his desktop—while handling the customer. The associate can be either physically on the BPO premises connected to the LAN (in-centre) or work from home (WFM) which allows to work at-home in any country. The main purpose of this study is to compare the recording effectiveness between associates operating in the contact centre premises (in-centre) verus associates working from home. The experiments show that there is a difference in the percentage of recordings achieved for in-centre associates verus the recordings carried out for work-from-home associates. The results clearly point out the need to bring about architectural changes to improve the effectiveness in the recordings at both types of working places. This paper also undertakes the experiment to verify if rebooting or restarting a user desktop at the end of each shift can impact recording percentages in a positive manner.

Keywords Screen recording · Performance · Compliance · BPO

R. Deshpande (✉)
Intelenet Global Services. Mindspace, Malad (West), Mumbai, India
e-mail: deshpande.rajendra@gmail.com

U. Shiurkar
Devogiri Institute of Engineering & Management Studies, Aurangabad, India
e-mail: shiurkar@gmail.com

S. Devane
Datta Meghe College of Engineering, Mumbai, India
e-mail: srdevane@yahoo.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_51

1 Introduction

Recording plays a vital role in measuring compliance, provides supporting information during disputes and plays a role in quality improvement to identify training needs.

India is one of the top destinations favoured globally for outsourcing as it provides 24×7 services bundled with fluent English speaking skilled labour along with strong political stability in the region, this has led to a tremendous growth of the BPO industry in India. Indian BPO companies offer varied services, such as Customer support services (Voice as well as data), Technical support services, Telemarketing services, Data entry and processing, etc.

One of the growing concerns in the outsourcing industry, amongst them the business circles and security and compliance community, was the need to monitor and eliminate the possibilities of compliance breach, especially in the financial industry, along with the need to meet the regulatory compliance in the event of dispute resolution and/or settle an arbitrary claim. The emergence of 100% screen and voice logging and recording are results of these growing concerns. Screen and voice logging and recording provide an end-to-end visibility of the operational activity. For this, transactional analysis and multichannel monitoring are required. Screen recording is a mechanism through which the workstation screen is recorded or captured during the interaction with the customer. The recordings are used to monitor advisor screen activity during the interaction to analyse their work efficiency and integrity. Both, the screen recording accompanied by the voice recording, are network agnostic, which means the recordings of desktops and voice are deployed either on LAN or remote offices connected over WAN.

Here, we describe the setup of the recording architecture and discuss the effectiveness and usage of the screen and voice logging for in-centre and at-home associates/agents from data derived from thousands of agents of both types of working places.

2 Literature Review

Business process outsourcing (BPO) is an emerging industry which requires a manifold of systems to assure quality and reliability but also safety and security for both clients and customers. For this, many BPO businesses applied screen and voice recordings to monitor the operational activity. We followed the approach of identifying practical issues that are faced in industry when the screen recording is deployed in the BPO environment. We have not seen similar study being carried out in this space though there are numerous researchers and companies that already provides some reference work in this context This was not touched effectively by the studies conducted by earlier researchers. We started our BPO specific environment research with the first 3 papers listed below. The prior research is listed post that.

Reference [1] R. Deshpande et al. (2017) [CDAC242, IJETMAS, June 2017] provided a model to predict the outcome of whether a call received at a contact centre is likely to be recorded or not.

Reference [2] R. Deshpande et al. (2017) [IJETAE, Oct 2017] defined a framework to ensure that screen and voice recording in a federated technology environment at a BPO occurs without any discrepancy.

Reference [3] R. Deshpande et al. (2015) provided the analysis of issues in screen recordings for the high-transaction BPO industry. This research addressed the following areas in the field of screen recording: (a) Identified the gap in current measurement processes and provided a clear approach for actual measurement processes and (b) Carried out a systematic identification of key issues responsible for not meeting a compliance.

Reference [4] Arockiasamy and Abdullah (2013) defined the service quality as a determinant of customer satisfaction. In most of the organizations, service quality derives from the service outcome from service providers. The authors pointed out that the perceived service quality acts as a component of customer satisfaction. In addition, service quality has the tendency to focus on evaluation and to reflect the perceptions of the customer with reliability, assurance, empathy, tangibility, and responsiveness.

Reference [5] Agarwal et al. (2013) discussed the impact of dimensions in quality service of call centres in India towards the satisfaction of the customers. The organizations are based on the stakeholders, i.e., the customers, so the satisfaction level is important and renders the service quality in the competitive environment of BPO. The service quality is essential for reducing the cost, increasing the profitability, raising the market share, satisfying the customers, and providing better competitive advantages. The call centres in India are found to give more importance to customers as they are the path to success, they interact with customers, deliver the better services, and enhance the technological performances.

Reference [6] Vidicode (2015) studied the recording of calls which are the most integral part of business developments and also provide security strategies for the audio analysis in any of the call services. These call recordings can provide reliability in the recording platforms of the call services with extra features for a better functionality and provide ideal solutions for the customers. The logs of the recording are managed for future use and can be played back and accessed anytime and every time with privacy and interiority from the managing systems. It acts as a valuable tool for protecting the data, i.e. the comments and feedbacks given by the clients, and it also makes screen recordings during each call, adds notes, start, stop, and tags the calls with its information.

Reference [7] Brent et al. (2001) conducted a study to improve the computer-based test and understand the impact of the screen resolution. This paper evaluated the effects of variations in screen size, screen resolution, and presentation delay.

Reference [8] Wu et al. (2014) developed a fast skip mechanism namely FMS (fast media synchronization) for matching the playback of the video stream by adjusting remote stream in the screen data. They identified that the mechanism of FMS approximately minimizes 85% of the time for processing, 57–99% for pasting time frequency, 70–90% of the amount for comparisons, and 70–90% for total pasting area.

Reference [9] According to the research by Kwou et al. (2014), a high-speed encoder in the video stream for standard specification of H.264 digital video codec is accelerated. On the basis of the techniques of parallel processing with GPUs (Graphical Processing Units), it adopted a programme of Open CL-based GPU kernel and finally it obtained a high-level CPU-GPU interoperability. That was a remarkable speedup compared to prior implementations.

Reference [10] Sanchez et al. (2013) adapted bi-directional and hierarchical inter-prediction on a GPU-based platform for 3D and 2D H.264 video coding. Standard of H.264/AVC video coding pioneers had some enhanced tools to maximize compression efficiency. With the introduction of many accelerators or core processors, a step was taken for deploying an H.264/MVC and H.264/AVC inter-prediction algorithm on GPU. The negligible distortion drop rate with a reduction time of up to 98% provided for full H.264/AVC encoder.

In summary, the different publications on screen and voice recordings as well as the techniques required for high-quality output of these recordings demonstrate that this topic still requires a lot of more research to understand the usage, transmission, and effectiveness of the screen and voice recordings and loggings in the BPO business. Our aim is to address in the following the required architecture for both the desktops of in-centre and at-home agents as well as to point out to some practical issues in their usage.

3 Recording Architecture in a BPO Enterprise

The recording architecture in a BPO enterprise requires several system to integrate both, the recordings from agents working locally at the centre and the agents working from home. Figure 1 shows the schematic architecture for screen and voice recording in the BPO industry along with the components involved. In this case, the in-centre agent (no. 7) is deployed on LAN and serial no. 14 represents an agent working from home using VPN. The architecture below also comprises an automatic call distributor (ACD, no. 2) and the screen and voice recording for in-centre as well as agents working from home.

Following steps shows how the voice and screen recording is achieved for in-centre agents.

1. Customer calls are routed through public switched telephony network (PSTN) network to the Automatic Call Distributor (ACD).
2. The ACD distributes the calls received from customer to an agent. Upon sending the call to the agent, the event information is generated by the ACD, which is passed onto computer telephony interface (CTI) interface.
3. The CTI transfers the event to the logger interface
4. A Local Area Network (LAN) allows interconnection of all components.

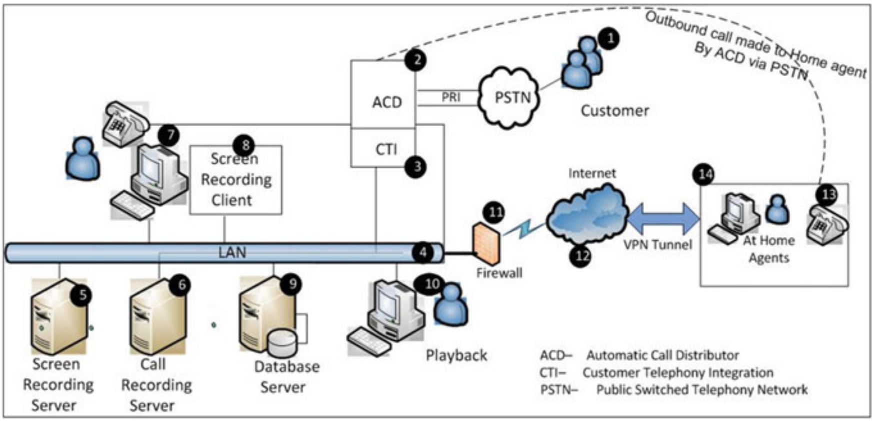


Fig. 1 Schematic architecture

5. The screen recording server records all screens and is also responsible for collecting and archiving the screen database.
6. The call recording server is responsible for collecting and archiving the voice calls.
7. Desktops are used by an agent for application usage during the conversation with customers
8. The screen recording client, is installed on each agent's desktop. It helps to identify the IP address of the agent workstation and sends the screen recording to the recorder.
9. The database server contains the entries of all calls recordings.
10. Play back is used for call listening, quality monitoring or for compliance monitoring purpose along with the screen play.
11. The firewall is used to build the necessary security and the VPN tunnel over the internet for the at-home agents.

The following steps show how the voice and screen recording is achieved for agents working from home and represents the components no. 11, 12, 13, and no. 14 in the architecture in Fig. 1.

- The at-home agent connects to the VPN-over-Internet while beginning the work.
- The call is routed to the home agent either via PSTN (the ACD routes the call using another channel using an outbound call) or VoIP over internet. In case of the at-home agent, since agents are not on LAN, active recording method i.e. service observe (barge in) or single step conference feature of ACD is used for voice recording. The screen recording agent is installed on the home agent desktop. When the home agent logs into ACD, his desktop gets also connected to the screen recorder.

Table 1 Data collection parameters

Sr. No	Parameters	Data collected points
1	Number of calls landed	Calls landed on the ACD (component 2)
2	Number of calls received	Calls received at the ACD (component 2)
3	Number of screen recordings	Screen recordings available on the recording server (component 5)
4	Number of voice recordings	Voice recordings available on the recording server (component 6)

- When a call is answered by the home agent, the voice recorder initiates a barge in or conference using the ACD resources and the voice is recorded with the voice recorder at the contact centre.
- On-call being received, the CTI event is read by the voice and screen recorder. The screen recorder initiates screen capture on the agent's desktop. The screen recordings are sent over VPN back to the main screen recording server.

4 Data Collection

The objective of the data collection was to analyse the screen recording in different scenarios without being bias during data collection. The study was based on the data collection carried for in-centre and at-home agents over a period of one month for the US geography. Data was collected by method of structured and controlled observation from the system. Following parameters were used for data collection and are clearly depicted in the Fig. 1.

The recording data was collected at the ACD for both types of agents (in-centre and at-home) as well as voice and screen recordings. The only difference between the environments for in-centre and at-home was the medium of access. In case of the in-centre agents the recording happened over LAN while the agents accessing the logger from home were using the internet over VPN. The data covered 69,436 calls taken by in-centre agents and 43,700 calls taken by at-home agents over the one month period.

5 Data Analysis

Table 1 shows the data collection sample and Fig. 2 depicts the screen recording compliance for in-centre and at-home agents (Table 2).

In case of the in-centre agents, the agent's desktop and loggers are on LAN. The IP phones are configured in Virtual LAN (VLAN) making sure that it has small broadcast domains. This limits this VLAN from having other noisy traffic from the

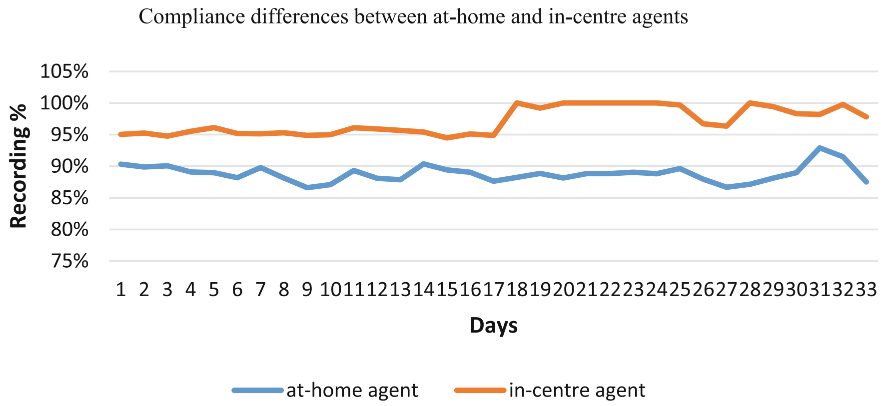


Fig. 2 Screen recording compliance of at-home and in-centre agent

LAN. This filtered traffic is then captured using Layer 2 sniffing technique. This technique helps to capture the RTP (real time packets) traffic from the Voice VLAN and diverts it towards the receiving LAN port configured on the logger. The logger holds the captured traffic and then transfers it to centralized archival depending on the retention rules set for the system.

The probability of losing a CTI event is very low in case of LAN usage for voice and screen recording. In LAN all the access ports communicate at 100/1000 Mbps which offers a huge throughput for Voice/Screen recording systems. Since all the ports are communicating at 100/1000 Mbps the latency is less than 10 ms with almost no jitter or packet loss. The Mean Opinion Score on the LAN is above 4.0 which makes sure that the original voice at source is of high quality.

Instead, the at-home agents use the Internet connectivity as the medium of communication between ACD and the home phone. The client–site VPN-over-internet is used for establishing the connection between ACD and the at-home agent's phone. The use- specific information like PSTN number used by the at-home agent is transferred to the ACD. The screen recording client installed on the at-home agent captures the user screen and sends the traffic to a centralized logger installed in the data centre. This traffic is transported over the VPN connection. The success of screen and audio recording depends on the captures done at the at-home agent's desktop. Since this traffic is transported over VPN, the captures are not 100% due to internet connection or VPN outages, flaps, network congestion, and lower quality of services of the links. An internet connection typically goes through multiple hops increasing the probability of packet drops at each hop. This reduces the percentage of screen and audio recording.

The similar exercise was carried by change of connectivity for remote agents at other office/location over the MPLS instead of internet while the in-centre agents were on LAN. Figure 3 provide the details of the increase in the percentage.

Table 2 Data Collection Sample

Date	In-centre/At-home agent	Calls taken (aggregate)	No. of screen recordings (aggregate)	Percentage screen recording compliance
01-Nov-17	In-centre	3730	3545	0.95
02-Nov-17	In-centre	3972	3783	0.95
03-Nov-17	In-centre	3875	3672	0.95
04-Nov-17	In-centre	4793	4579	0.96
05-Nov-17	In-centre	3511	3374	0.96
06-Nov-17	In-centre	2737	2605	0.95
.				
.				
14-Nov-17	In-centre	443	435	0.98
15-Nov-17	In-centre	449	449	1.00
16-Nov-17	In-centre	503	492	0.98
.				
.				
.				
01-Nov-17	At-home	1546	1364	0.88
02-Nov-17	At-home	1501	1334	0.89
03-Nov-17	At-home	1500	1322	0.88
04-Nov-17	At-home	824	732	0.89
05-Nov-17	At-home	699	621	0.89
.				
.				
06-Nov-17	At-home	1442	1284	0.89
14-Nov-17	At-home	1848	1717	0.93
15-Nov-17	At-home	1541	1410	0.91
16-Nov-17	At-home	496	434	0.88
.				
.				

When compared with the previous methodology (Fig. 2) we see a significant increase of around 9% by a mere shift if in the connectivity from Internet over VPN to MPLS (Fig. 3).

6 Conclusion and Future Work

During the experimental analysis conducted on BPO screen recordings the following observations were made

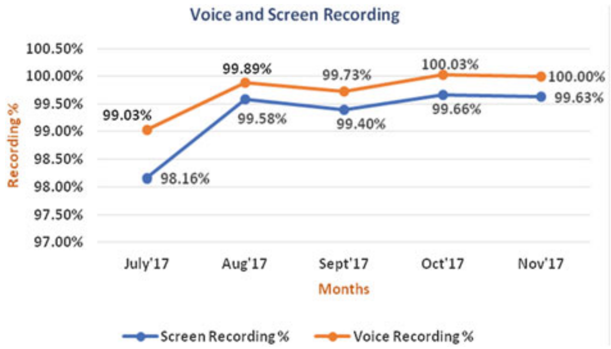


Fig. 3 Voice and screen recording percentage increase in the subsequent months

- The analysis shows that there is an 9% increase when we shift the connectivity from Internet to MPLS for the at-home agents.
- The analysis shown in Fig. 2 concludes that there is an 7% drop in the percentage of screen recordings when comparing in-centre agents and at-home agents.
- The loss of recordings is mainly due to the Internet connectivity which, when compared to LAN, has more latency, jitter, as well as packet drops.
- The reduction for at-home agents also happens due to the recording initiation trigger not reaching the desktop in the VPN environment.
- Future work can be performed to study the impact of latency on the recordings. Deployment of an SD-WAN (software-defined wide area network)

Note

Appropriate permission for the responsible authorities has being taken for data collection and analysis.

References

1. Deshpande, R., Shiurkar, U., Devane, S.: Provides model for predicting call recording at a contact center. [CDAC242, IJETMAS, June 2017, ISSN 2349-4476]
2. Deshpande, R., Shiurkar, U., Devane, S.: Analysis on lack of closed loop feedback in recording and the proposed framework. [IJETAE/ Certificate/ Volume7/ Issue10/ 37, Oct 2017, ISSN 2250-2459]
3. Rajendra Deshpande, Ulhas Shiurkar, Satish Devane (2015). “Analysis of issues in screen recording for high transaction BPO Industry”[International Joint Colloquiums on Computer Electronics Electrical Mechanical and Civil, Grenze ID:02.CEMC.2015.1.15]
4. Arokiasamyy, A.R.A., Abdullah, A.G.: Service quality and customer satisfaction in the cellular telecommunication service provider in Malaysia. J. Arts Sci. Commer. **4**(2) (2013)
5. Agarwal, S., et al.: Impact of service quality dimensions towards customer satisfaction in indian call centres. Pac. Bus. Rev. Int. **6**(1), 51–64 (2013)
6. Vidicode, (2015), Cost Effective Call Recordings Without Compromise, Apresa, pp. 1–7

7. Brent, B., Mary, L.L., Altamese, J.: Effects of screen size, screen resolution, and display rate on computer based test performance, Research report by ETS, Statistics and Research Division, Princeton, Oct 2001, RR-01-23
8. Wu. P., et al.: A fast media synchronization mechanism for streaming screen recording applications. In: IEEE 17th International Conference on Computation Science and Engineering (2014)
9. Kwou, Y., et al.: A parallelized implementation for H.264 real-time encoding scheme. Int. J. Control Autom. **7**(6), 379–388 (2014)
10. Sanchez, R., et al.: Adapting hierarchical bidirectional inter prediction on a GPU-based platform for 2D and 3D H.264 video coding. EURASIP J. Adv. Signal Process. **67** (2013)

Document Theme Extraction Using Named-Entity Recognition

Deepali Nagrale, Vaibhav Khatavkar and Parag Kulkarni

Abstract The text mining can be implemented by term analysis of word or phrase. This term which describes the concepts of particular sentence is use to define the document theme. The new context-based mining technique is introduced which uses the concept-based mining model to analyze the terms present in sentence, document, and corpus levels. We find the theme of document like organization, medical, entertainment, sport, and so on. Context-based mining apply on statistical data as well as real-time data like Export data from Wikipedia. The theme of document is extracted by using Natural Language processing (NLP) for communication between computer and human languages and name entity recognition (NER) algorithm for identification of entity, entity chunking, and entity extraction. It used to get name entity in text such as person name, organization name, specific locations, time expressions, percentages quantities and so on. NLP and NER are used in context-based mining for finding name of entity and their relationship. Context Vector containing set of documents is used to extract the context of the document. Finally K-Mean algorithm is used for clustering to find inherent groupings of the text documents, then set of clusters are generated where each cluster exhibit high intra cluster similarity and low inter cluster similarity. The text document clustering is used to separate documents into groups or clusters based on their similarity so all groups define the distinct topics.

Keywords Context mining · NLP · NER · Context vector

D. Nagrale (✉) · V. Khatavkar · P. Kulkarni
Department of Computer Engineering and IT, College of Engineering, Pune 411005, MH, India
e-mail: nagraledy16.comp@coep.ac.in

V. Khatavkar
e-mail: vkk.comp@coep.ac.in

P. Kulkarni
e-mail: paragindia@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_52

1 Introduction

The text in the document is represented in the form of characters, words, and sentences. The Context of the text document is the abstract representation or any information that describes the document. Context is available in various forms of the mining model which is based on sentence-, document-, and corpus-based concept analysis and concept-based similarity measure [1]. We used NER, NLP, and K-mean algorithm for context mining, NER algorithm is used to get named entities of text such as person name, organization, locations, expression of times and so on [2]. K-Mean algorithm is used for clustering to identify inherent groupings of the text documents and gives results by sentence-, document-, and corpus-based. Here similarity measure considering concept is evaluated by considering the important concepts with respect to semantics of sentence, description of document, and differences among all documents. Therefore instead of using single term analysis measure, combined approach for concept analysis measure is used to gives best quality. Here we used an unsupervised learning method which is used to find inherent groupings of the text documents, then set of clusters are generated where each clusters exhibit high intra cluster similarity and low inter-cluster similarity. The text document clustering is used to separate text documents into no of groups or clusters based on their similarity so all groups define the distinct topics.

Many researchers has worked on document processing for extracting the context of document/s that includes various methods like latent sentiment analysis, context vectors, Association Mining, joint sentiment/topic model, etc [3–6]. The context of a document can be found by generating vector space which is the combination of set of documents using various machine learning algorithms but they fail to give best accuracy for large datasets. This can be handled by using Latent Semantic Analysis (LSA) [4]. Our proposed system parsed the long documents, tokenized each document, and extracted the context from text document.

Section 2 describes literature survey, the work done on extracting context of a document and related work, Sect. 3 proposed a context extraction system and Sect. 4 describes experimentation and results. And finally conclusion and future work defined in Sect. 5.

2 Literature Survey

We consider text documents in large to find the relation between different documents in corpus. Hugh document collection with long documents are analyzed so that it become difficult to derive insights by looking only for representative concepts in the selected document collection based on a divergence metric. We use Natural language processing (NLP) and Natural Language Toolkit (NLTK) a Python library that provides modules to process, classify, tokenise, stems, tag, parse, and wrap text [7].

Kundan et al. [3] proposed a system which created a dictionary to store different contexts and associated terms of text documents which considered the context of document and context of user separately so it avoid delivery of documents to unintended user over the shared network but the tests were performed on sets of documents which were not standard.

Katja et al. [8] proposed a system in which learning objects by analyzing their use in web portal are used to detect semantic similarity measures. Here usage-based relations between the objects are considered rather than considering the content of learning objects or user object relationship. This new approach is used to enhance existing approaches in technology enhanced learning.

Dharmendra et al. [9] define measure similarity between concept vector and context vector by using Vector Space Model (VSM). Google APIs were used to extract data from various web applications like news site, twitter etc. According to the authors, their system had low calculation cost than other existing systems as it removes sparse terms during the process of context extraction.

Yifei et al. [10] used feature selection methods to define four new context similarity. The model was developed using Support Vector Machine (SVM). The testing was done on two standard datasets namely DataBCII and Reuters-21578. Methods were analyzed and compared against frequency based techniques. Author defined to use semi-supervised method instead of SVM in future work.

Mayanka et al. [11] proposed the VisContext framework for context learning from large-scale multimedia data. The implementation of the VisContext framework was presented and framework was evaluated on large datasets to find the accuracy in terms of visual understanding. The results are calculated based on contextual association of terms and images, clustering and classification of images based on context.

Kulkarni et al. [12], also proposed a system to extract context based on Naive Bayes and Apriori Association. The tests were performed on abstracts of 57 IEEE papers but there was lack of standard dataset for testing.

Ankit et al. [13] proposed a method of E-VSM in which text documents are used to perform density based clustering. It performed ranking of target documents but there was lack of standard datasets for testing.

Asef et al. [14] in their work used Latent Semantic Analysis (LSA) to extract Context in persian language using multidocument summarization. Author says the proposed system gives better performance than existing persian multidocument systems.

Vaibhav et al. [4] in their paper, worked on a "Context Vector Machine" to define the theme of a document and extension work to identify the "Context Vector" of a document set is done in the paper [5].

Wahiba et al. [23], suggested learning approach to recognized Named-Entity problem by identifying and classifying the contexts, those which are frequently found with the NE. Contexts means the words present left or right of the NE was used to find NE which has asigned to a NE class. Training corpus: web documents is used. Context weights associated to context frequency, document frequency in the corpus is calculated by frequency and modified tf-idf representations.

3 Proposed System

We are using Named-entity recognition (NER) algorithm to extract the context from plain text sentences of the document, which store and sort textual content into default categories such as person name (PER), organization name (ORG), specific locations (LOC), time expressions, percentage, quantities and so on [1, 2]. The sequences where these words were found are then label and store in module for further analysis.

For example, the following table shows a simple input sentence, and the terms and values generated by the module:

Input text	Module output
"Boston is a great place to live."	0, Boston, 0, 6, LOC

The output can be interpreted as follows:

- The first "0" means that this string is the first article input to the module. Because a single article can have multiple entities, including the article row number in the output is important for mapping features to articles.
- "Boston" is the recognized entity.
- The "0" that follows "Boston" means the entity "Boston" starts from the first letter of the input string. (Indices are zero-based.)
- "6" means the length of "Boston" is 6.
- "LOC" means the recognized "Boston" is a location. Other supported named-entity types are person (PER) and organization (ORG) (Fig. 1).

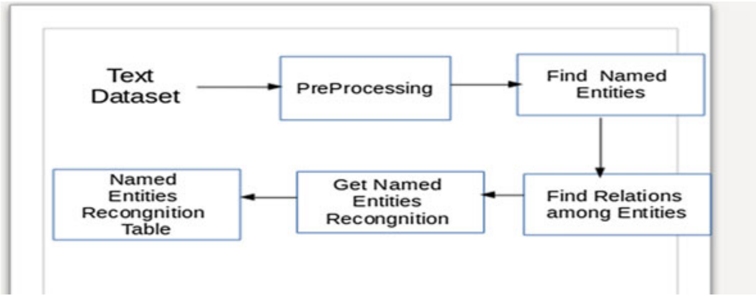


Fig. 1 Named-entity recognition (NER) proposed system

3.1 Context Extraction Algorithm

1. Take text dataset as input to the system.
2. Pre-processed the data.
3. Tokenized the document.
4. Removed stopped words and punctuations from each document and find named entities.
5. Find the relations among the named entities generated from step 3.
6. Get named-entity recognition (NER) to find the context of a document.
7. Generate named-entity recognition (NER) table.

Existing System Mechanism

The proposed model differentiates the documents having same similarity measures. Large sets of experiments are conducted on mining model which is based on concept on large data sets using text clustering. The experiments differentiate the traditional analysis and concept-based analysis (Fig. 2).

Proposed System Architecture

The proposed a system is used to extract important information from number of large documents. K-mean algorithm is used to cluster similar data from statistical data as well as real-time data. We find the sentiment between documents using NER, NLP and result is used to cluster the documents (Fig. 3).

Let E is a set of entities in documents.

Let A be the set of all values in E

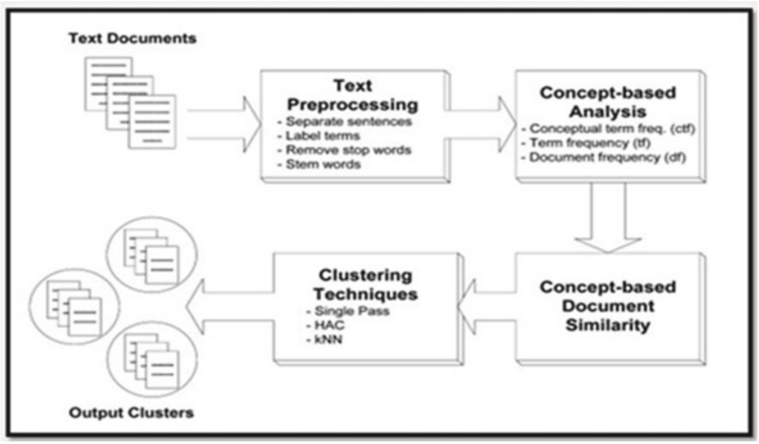


Fig. 2 Existing system mechanism

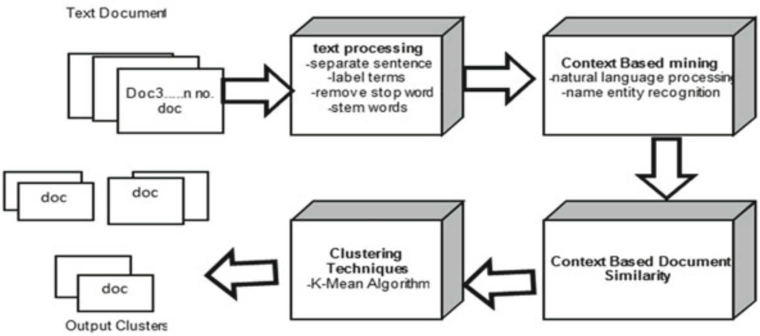


Fig. 3 Proposed system flow

Entity set for NER Class-3

E1={Location, Other, Organization, person}
A = {e ∀ E1}
= {{Briten,UK, US, Andreas, Europe, Japan, Boston, Heringsdorf, Stuttgart},
{European, company, production, publication, shares, stock},
{Bank, Forbes, Royal, Scotland, Vodafone, Microsoft, Research, Parliament, Scottish, University, Southampton, BBC},
{Chris, Ferguson, Lyle, David, Yarnton, Adrian, Alan, Howard, Matthew, Robert, Russell,Taylor}
}

Theme of a document = Max{e} where e is the cardinality of a document.

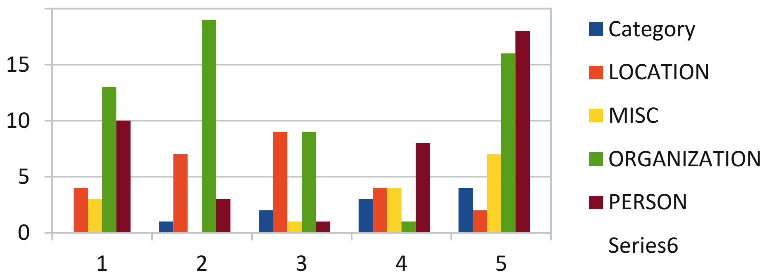


Fig. 4 NER Class4 output

Entity set for NER Class-4

E2={Location, Misc, Other, Organization, Person}
A = {e ∀ E2}
= {{Boston, Heringsdorf, Stuttgart, Briten,UK, US, Andreas, Europe, Japan,},
{FT-30, FTSE, Scottish, British, Country, Olympics, BBC, Catholicism, News, Tory},
{European, company, production, publication, shares, stock},
{Microsoft, Research, Parliament, Bank, Forbes, Royal, Scotland, Vodafone, Scottish, University, Southampton},
{Howard, Matthew, Robert, Chris, Ferguson, Lyle, David, Yarnton, Adrian, Alan, Russell, Taylor}
}
Theme of a document = Max{e} where e is the cardinality of a document.

Entity set for NER Class-7

E3={Date, Location, Other, Organization, Percent, Person, Time}
A = {e ∀ E3}
= {{ 1935, 2003, 2005, December, January, March, Sunday, Mid-November, Friday, Week},
{ San, Briten,UK, Boston ,US, Andreas, Europe, Japan, Heringsdorf, Stuttgart},
{Company, Consmer, FT-30,Conference, FTSE, Scottish, British, Country, Olympics, BBC, Catholicism, News, Tory}
{European, company, production, publication, shares, stock},
{Scotland, Vodafone,Bank, Forbes, Royal, Microsoft, Research, Parliament, Scottish, University, Southampton},
{55},
{Tanguay, Jo, Chris, Ferguson, Lyle, David, Yarnton, Adrian, Alan, Howard, Matthew, Robert, Russell, Taylor},
{afternoon, evening}
}

Results for NER Class 4, NER Class 3 and NER Class 7 are explained and shown in Figs. 4, 5 and 6 respectively in experimentation section.

In above Architecture, NLP and NER are used to find name entity relationship, similarity and sentiment between documents. In final stage clustering on similar documents is done by using k-mean clustering algorithm as follows

1. Consider “k” as a no of Clusters. And Randomly select the k cluster centers.
2. Calculate mean or center of the cluster and Calculate the distance between each point to each cluster center.

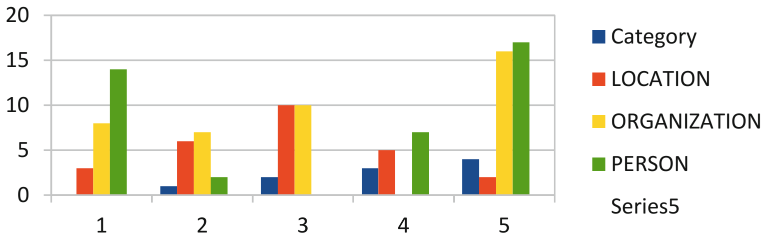


Fig. 5 NER Class3 output

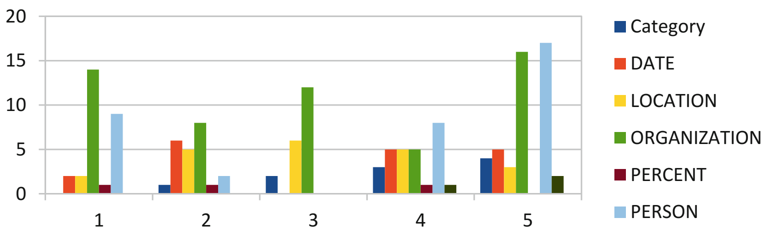


Fig. 6 NER Class7 output

3. If the distance is closer to the center then move to that cluster otherwise move to next cluster.
4. Re-establish the center.
5. Repeat the process until the center does not move.

Cluster means is calculated as follows:

$$Mean = \frac{\sum_{i:c(i)=k} Yi}{Nk}, \mathrm{k} = 1, 2, \ldots, \mathrm{K}.$$

The Distance between the each data point and cluster center is calculated as

$$\mathrm{D(i) = average\ of\ min} \|\mathrm{Y_i - Mean_k}\|^2, \mathrm{i} = 1, 2, \ldots, \mathrm{N}.$$

Repeat the above two steps until mean value convergence.

4 Experimentation and Results

Standard BBC news dataset is used for experimentation. BBC news dataset consists of five category of news namely “business”, “entertainment”, “politics”, “sports”, and “tech” [15]. There are total 2,225 news in the form of text files. All these text files from dataset were given as input to the program. The proposed system tested ones NER Class 3, Class 4 and Class 7. Five category of news are taken on x-axis and all values of extracted entities from each document in dataset is taken on y-axis. From set of entities and their values, we calculated the cardinality for each set and depending on the maximum cardinality, the theme of a document is defined.

The Output for NER Class 4 is shown in Fig. 4, here the cardinality for each entity is calculated and maximum cardinality is used to define the theme for particular news category. E.g. “Organization” entity is having maximum cardinality as compare to other in “business” and “entertainment” category of news so we can define the theme of “business” and “entertainment” categories of news as “Organization”, in “politics” category of news, cardinality of “Location” and “organization” entity is equal and high so we can define the theme of “politics” category of news as “Organization and Location” both, similarly we can define the theme for “sports” and “tech” category of news as “Person” as it having maximize cardinality as compared to other and so on. The same results for NER Class 3 and NER Class 7 are shown in Figs. 5 and 6 respectively.

The Accuracy Comparison for NER Class 4, Class 3 and Class 7 is shown in Table 1. By considering the maximum cardinality the theme for particular category of news is defined and compared for different NER Classes. Therefore “Organization” entity which is having the maximum cardinality for “Business” news as 13, 8 and 14 for NER Class 4, Class 3 and Class 7 respectively. After comparing cardinality for different NER classes, NER Class 7 gives maximum accuracy as compared to NER Class 3 and Class 4. Similarly for second category “Entertainment” news, NER Class 4 gives maximum accuracy for “Organization” as 19 as compared to NER Class 3 and Class 7 for “Organization” entity which gives accuracy as 7 and 8 respectively. For third category “Politics” news, NER Class 7 gives maximum accuracy for “Organization” as 12 as compared to NER Class 4 and NER Class 3 which gives maximum cardinality for “Organization and Location” as 9 and 10. Similarly for forth category “Sports” news, both NER Class 4 and Class 7 gives equal and maximum accuracy as compared to NER Class 3 and for fifth category “Tech” news, NER Class 3 and Class 7 gives best accuracy than NER Class 4.

5 Conclusion and Future Work

As compared to already existing systems from literature survey, this is an effective model. Context-based mining model proposed to find theme of the document from the extracted named entities such as the name of person, organization, location, expression of times, quantities, monetary value, percentage, and so on. The proposed system is tested for different NER Classes on same dataset for accuracy comparison and then K-Mean algorithm for clustering on similar data is used to get inherent groupings of the text documents. Future work will be testing the proposed system on different datasets and result comparison and implementation of LSA or PLSI on given BBC dataset and compare this results with our proposed system.

Table 1 Comparison of results of NER Class 4, Class 3 and Class 7

	Business		Entertainment		Politics		Sports		Tech	
	Highest cardinality term	Max Value	Highest cardinality term	Max Value	Highest cardinality term	Max Value	Highest cardinality term	Max Value	Highest cardinality term	Max Value
NER Class 4	Organization	13	Organization	19	Organization and Location	9	Person	8	Person	18
NER Class 3	Organization	8	Organization	7	Organization and Location	10	Person	7	Person	17
NER Class 7	Organization	14	Organization	8	Organization	12	Person	8	Person	17

References

1. Kamath, S., Wagh, R.: Named entity recognition approaches and challenges. Int. J. Adv. Res. Comput. Commun. Eng. ISO 3297:2007 Certified 6(**2**) (2017)
2. Sharnagat, R.: Named Entity Recognition: A Literature Survey (2014)
3. Dhande, K.A., Umale, J.S., Kulkarni, P.A.: Context based text document sharing system using association rule mining. In: Annual IEEE India Conference (INDICON) (2014)
4. Khatavkar, V., Kulkarni, P.: Document context indentification using latent semantic analysis. In: Presented in: 3rd International Conference On Computing, Communication, Control And Automation, 17–18 Aug 2017, Pune, MS, India (To be published on IEEE)
5. Khatavkar, V., Kulkarni, P.: Context Vector Machine for Information Retrieval. Atlantis Press, Advances in Intelligent Systems Research **137**, 375–379 (2017)
6. Sowmiya, J.S., Chandrakala, S.: Joint sentiment/topic extraction from text. In: IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT) (2014)
7. https://pythonprogramming.net/named-entity-recognition-stanford-ner-tagger
8. Niemann, K., Wolpers, M.: Creating usage context-based object similarities to boost recommender systems in technology enhanced learning. J. Latex Class Files **6**(1) (2007)
9. Sharma, D., Jain, S.: Context-based weighting for vector space model to evaluate the relation between concept and context in information storage and retrieval system. In: International Conference on Computer, Communication and Control (IC4) (2015)
10. Chen, Y., Han, B., Hou, P.: New feature selection methods based on context similarity for text categorization. In: 11th International Conference Fuzzy Systems and Knowledge Discovery (FSKD),, 2014
11. Chandrashekar, M., Lee, Y.: Visual context learning with big data analytics. In: IEEE 16th International Conference on Data Mining Workshops (2016)
12. Kulkarni, A.R., Tokekar, V., Kulkarni, P.: Identifying context of text documents using Naïve Bayes classification and Apriori association rule mining. In: 2012 CSI Sixth International Conference on Software Engineering (CONSEG) pp. 1–4, 5–7 Sept 2012. https://doi.org/10.1109/conseg.2012.6349477
13. Bhakkad, A., Dharmadhikari, S.C., Emmanuel, M., Kulkarni, P.: E-VSM: novel text representation model to capture contex-based closeness between two text documents. In: 7th International Conference on Intelligent Systems and Control (ISCO) (2013)
14. Poormasoomi, A., Kahani, M., Yazdiand, S.V., Kamyar, H.: Context-based persian multidocument summa- rization (global view). In: International Conference on Asian Language Processing (2011)
15. BBC news dataset. http://mlg.ucd.ie/datasets/bbc.html

AnaData: A Novel Approach for Data Analytics Using Random Forest Tree and SVM

Bali Devi, Sarvesh Kumar, Anuradha and Venkatesh Gauri Shankar

Abstract Big Data has been coined to refer different types of automated and non-automated system, which generated huge amount of data like audio, video, PDF documents, medical, biometric, etc., in the form of structured, unstructured or semi-structured data. In this paper, we are representing data analytics using Random Forest Tree and SVM (Support Vector Machine). The Big Data Analytics is utilized after integrating with digital capabilities of business or other. As per our novel algorithm approach, we have modified a combination of two robust algorithms of data mining such as Random Forest Tree and SVM. To check the robustness and feasibility of our approach, we are using some statistical techniques like precision, recall, sensitivity, specificity and confusion matrix for proving accuracy and ability benchmark. At last, the accuracy and speed-up time for doing the analysis is low as compared to existing algorithm. As for the accuracy calculation, our approach 'AnaData' gives result as 95% approximately.

Keywords Big data analytics · Big data · Random forest tree · Support vector machine · Data mining

1 Introduction

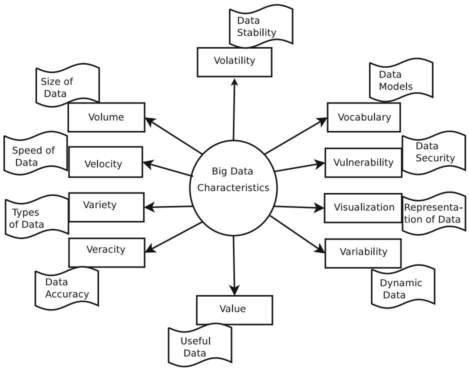
Big Data decodes the previously untouched heterogeneous mix of structured and unstructured data. This data comes from audio, video, PDF documents, medical records (such as X-rays), manuals, images, emails attachments etc., as the form of unstructured data and rows, column in DBMS tables as the form of structured data. Typical landscapes of big data include a huge data amount, which may be in

B. Devi (✉) · S. Kumar · Anuradha
CSE, Jayoti Vidyapeeth Women's University, Jaipur, India
e-mail: baligupta03@gmail.com

V. G. Shankar
SCIT, Manipal University Jaipur, Jaipur, India
e-mail: venkateshgaurishankar@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_53

Fig. 1 Big data characteristics



numerous data types, high requirement for processing speed and high analysis value return. The demands for huge data analytics are basically concentrated in several categories, like classifier, association rules and clustering [1, 2]. As a common method of data mining, Random Forest method has been proved to be a state-of-the-art of learning model, which having robust classification, regression with fast and efficient operations. Random forest method is not only the subject to memory limitations and featured with quick processing speed and good parallel scalability, but also an exceptional classification tool which can handle to gigantic data and an archetypal decision tree classification algorithm. We can see big data in action everywhere nowadays, and also the collection of datasets mainly follow the concept of 10 V^s [3, 4].

1.1 Big Data Characteristics

The characteristic of big data is based on the 10 V^s—Volume, Velocity, Variety, Veracity, Variability, Validity, Vulnerability, Volatility, Visualization and Value. The description of above 10 V^s are given below [5] as shown in Fig. 1.

- **Volume**: This is the collection of huge amount of structured, unstructured or semi-structured data which are stored in enterprise repositories. It explains to the vast collection of data which is created every second from social network, smartphones, transactional card, sensors, images, etc. We can produce and share this type of data each and every second. We can use heterogeneous framework, where segments of the data is filled in many positions and link created by software [6].
- **Velocity**: This describes the rapidity of data processing. Velocity is feasible when incoming data is slower in rate. We can take an example like social networking sites Facebook, Twitter, LinkedIn, etc. In Facebook, it generates 2.9 billion-like actions per day and 450 million photos. On other hand, an eminent example of velocity as Google processes over than 1.5 trillion searches per year.

- **Variety**: Data comes in different types of formats. In structured data like numeric, databases (in table forms), whereas unstructured data like video, audio, financial transaction, text documents, etc. Unstructured data comes from social networking sites or satellites. In survey, 80% of the world's data is now unstructured as of photos, video arrangements or social networks. Very New and advanced big data analysis is now permitting unstructured and structured data to be collected, stowed, and used concurrently.
- **Veracity**: It refers to the reliability or trusty of the data. With countless values of big data, quality and correctness are less manageable (As think with Twitter posts with hashtags, mistakes and informal text as well as the consistency and correctness of content) but Big Data analytics process now permits us to research with these variety of data. We can take some other analogy of this that links with the value of GPS oriented and satellite data are lost as they combine the tall tower or other instance. As soon as this executes, setting of data, which is attached with other data origin like data extract with accelerometer or road data to provide right data [7].
- **Value**: There is one other 'V' with improved version when looking at Big Data, i.e. Value. Value with correctness having rights to big data but unless we can change it's value, it is rough. In survey, the data created within the last 2 years in the world is about 90%. Data is being produced at astronomical rates and big data is a continuously modified factor with latest tools are daily being created to handle. Value is significant that enterprises make a business scenario for any effort to fetch big data today.
- **Variability**: Variability in big data domain indicates to some variable things. Big Data often creates variable data sources and their domains. Framework created into organization can create especially, rough data. Big data is in different profiles because of the multitude data domains evaluating from many variable data domain and sources. Variability refers to the invalidate speed at which large data is created into user database.
- **Validity**: Validity is similar to veracity. Validity describes quality of data is for its intended use. The revenue from huge data analysis is only its passing data, so it is necessary to acquire robust data in handy to ensure best data quality, data definitions, and data about data.
- **Vulnerability**: Big data carries few security issues with data. After extracting data, big data is a big issue. Unfortunately, there are many big data breaches. For an analogy identified by CRN, an attacker appealed bland posted big data on the high dark web to deliver in May 2016, which assuredly linked data on 167 million LinkedIn profiles, 360 million emails and passwords for MySpace users' [8, 9].
- **Volatility**: At big data, associations indicate to store data uncertainly—a couple of terabytes of large data would not make high stockpiling costs; data is kept in the live server without importing processing issues. Because of the speed and volume of large data, its instability should be carefully considered.
- **Visualization**: Current big data perception confronts specialized difficulties because of constraints of in-memory innovation and poor adaptability, usefulness and reaction time. We require diverse methods for speaking to data. For example,

grouping or utilizing treemaps, parallel directions, round system graphs or cone trees [6].

1.2 Support Vector Machine: Introduction

Support Vector Machine indicates an algorithm based on the concept of decision planes for learning classification and regression rules from data that define decision boundaries. A decision plane define separation between a set of objects which have different class memberships. SVM can be used to understand polynomial equations, radial basis function (RBF) and multi-layer perceptron (MLP). SVMs were first developed by Vapnik in the 1960s. This algorithm is developed for classification and methods for theory coupled with extensions to regression and also known as density estimation. The concept used in SVM is principle of structural risk minimization which is closely related to theory of regularization. There are two key elements used in the implementation of SVM, which are the mathematical programming techniques and kernel functions techniques. We are defining SVM with the help of an example:

In this example, we have taken some objects which having two classes as class GREEN and RED. The dividing line between these objects define a border on the right edge of that having all GREEN objects and to the left edge of that having RED object. When we enter any new object into the left edge is tagged that is identified, as GREEN or vice versa.

The below example of a linear allotment or classifier, which divides a set of GREEN and RED objects into the relative groups with a line. This classification works based on visualizing different types of dividing lines to differentiate objects in different class member are also known as hyper-plane allotment or classifier. SVM is particularly suitable to handle such types of tasks.

The solution of this type of problems shown in Fig. 2 is the basic idea behind SVM. We can see here the original objects which mapped systematically. Here, they will use a set of logical methods which is defined as kernels. The flexibility of kernel functions allow the SVM to search a wide variety of hypothesis spaces. The routine of systematically rearranging these objects is defined as mapping or transformation. In this separation, instead of maintaining the complex curve as in left-hand side is schematic, we have to explore an optimal best line, which can be divide the RED and the GREEN objects [10]. The whole process of SVM is shown in Fig. 2.

Support Vector Machine: Limitations

- Limitation of SVM is speed and size in case of training as well as testing.
- Another problem with SVM is discrete data representation.

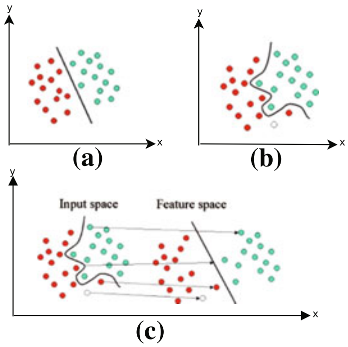


Fig. 2 Support vector machine: an example

1.3 Random Forests Tree

Random forest tree method is a classification as well as regression method proposed by Breiman in 2001. It uses bagging method. Bagging is to average noisy and unbiased models to create a model with low variance. RFT algorithm works as a large collection of de-correlated decision trees. Using bagging method, random forest method will draw multiple training sample sets that are different from each other. Every single sample set builds a decision tree with randomly selected attributes. Random forest is a collaborative learning (both classification and regression) technique. It is one of the frequently used analytical modelling and machine learning technique. RFT is developed for loss of decision tree based on random selection of data and variables [11, 12].

Random Forests Tree: Steps of Working

- Random selection of 'x' features out of total 'n' features. Where $x \ll n$.
- Out of the 'x' features, find the node 'k' using the best divide place.
- Arrange the node into daughter nodes using the best division.
- Repeat steps from 1 to 3 until 'd' number of nodes has been found.
- Maintain forest by repeating steps from 1 to 4 for 'm' number times to create 'm' number of trees.

Random Forests Tree: Limitations

- The strong limitation of the random forests tree algorithm is that a huge number of trees analysis may take more time to analysis, due to which process is slow for real-time prediction.
- This algorithm is not fit for some datasets with the task as noisy classification or regression.

2 Background

Raghavendra et al. [13] defined many big data issues in computing with robust and efficient ways of analytics. The big feature of this article is based on validation, visualization, verification and big data management. Yang et al. [7] identified the issues of cloud storage and its challenges on big data platform. This article manages the requirement processing capability and analogical skills in big data. This article also concentrates on digital earth or GPS domain with big data computing. Venkatesh et al. [14] presented a tool as 'Anti-Hijack', which finds session and intent issues using honeypot technique. They analysed big data of Android malware for vulnerability and security measurement. Stefan et al. [11] presented how to estimate standard errors for random forest. They also work with builds on variance estimates for bagging proposed by Efron (1992, 2013). Scornet et al. [12] proposed the consistency of random forests in big data. They have explained many good practices of random forest in many contexts. Brinker [10] explained incorporating variety in machine learning with support vector machine. They also explained computational complexity, statistical methods and computing for big data. They also define the role of computational statistic in scientific discovery from big data analysis recognize by peer statisticians. Breiman [15] demonstrated random forest in which, each tree trust on the data values of a random sampled vector unrelatedly and with the equal partition for all node trees in the forest. Bernard et al. [16] explained the introduction of a new Random Forest investiture algorithm that depends on a collective tree investiture process. The main aim is to instruct the tree investiture, because any tree is equal balance tree with the present tree.

Our approach 'AnaData' represents the robust approach to overcome the limitation of random forests tree and support vector machine. In the proposed methodology section, we will explain the fast accessing speed and variant size of training as well as testing dataset with homogeneous data representation. Our proposed algorithm also works with real-time system with fit for dataset with noisy classification.

3 Proposed Methodology, Algorithm and Dataset

We have the proposed framework which depends on data classification and analysis with novel approach as 'AnaData'. Our approach is based on random forests tree and support vector machine, which is a modified novel approach named as 'AnaData'. We have applied and executed our proposed framework on the huge dataset from social media, transactional, movies and historical, smartphone, geographical, GPS, genetic, website clickstream domain, etc. The steps of our proposed analysis 'AnaData' is given below:

- Fetch the dataset from the different repository.
- Place whole dataset ascending or descending way accordingly.
- Select the random data for finding the median on the basis quartile.

- Find the median (m1, m2, m3, m4) of dataset using quartile (q1, q2, q3, q4) measurement.
- Arrange the feature of data with median.
- Extract the best data value for median in each quartile.
- Combine the category of all median value with cluster.
- Create the vector category to find the nearest value of median.
- Predict the zero days on the basis of vector prediction category.
- With the help of modified algorithm of SVM and RFT, finally we have selected the random data and predict the data vector with the classification as Structured, Unstructured and Semi-structured.

We have done analysis on very large training set data and find better analysis outcome with SVM and RFT. In case of feature selection, we have processed it with median and find the predictive category with nearest median vector. The framework algorithm of 'AnaData' is represented in the Algorithm 1.

Algorithm 1 AnaData: Classification and Analysis

```
procedure DATASET,REPOSITORY(item d, r) Retrieve the dataset and repository
    for m = 1 → n do
        if q1 → Q4 == M then
            return to whole dataset
Check the Quartile
Check the median
Maintain the median vector
        end if
    end forFind p in (1 → n,q1 → Q4)    for i = 1 → x do
    if median_query(M, Q) == TRUE then
        y ← Q_{M%n}
        for item y ∈ d do
            if y.median == r then
                include y as nearest neighbor
Maintain and arrange all quartile
Finally get the median vector with classification
            end if
        end for
    end if
end for
end procedure
```

3.1 AnaData: Dataset Collection

We have implemented our proposed framework on the huge dataset from social media [17], transactional [18], movies and historical [19], smartphone [20], geographical [21], GPS [20], genetic [22] and click streamed [20]. We have analysed approach with accuracy of data with robust data classification in fast accessing as compared to existing algorithms and approaches in the same perspective using identical dataset. As for classification and analysis, our approach returns 95% of data retrieval and classification based on data types using collective dataset collection. See Table 1.

Table 1 AnaData: dataset collection

Dataset repository	Dataset type	Data size (Mb)
KDnuggets	Social media	330
UCI	Transactional	270
IMDB	Movies and historical	360
Google	Smartphone	256
Rtwilson	Geographical	310
Google earth engine	GPS	243
KDD	Genetic	156
Google	Click streamed	196

Table 2 AnaData: Result Statistics

	RFT		SVM		AnaData	
	Time	Accuracy (%)	Time	Accuracy (%)	Time	Accuracy (%)
Correctly classified	13776 ms	76.90	12139 ms	79.50	11230 ms	85
Incorrectly classified		23.10		20.50		15

4 AnaData: Result Statistics

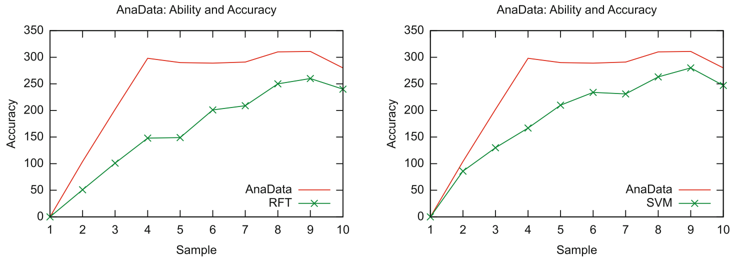
As per concern with result statistics of 'AnaData', we have improved RFT and SVM algorithm with our modified algorithm. Our framework provides fast and accurate data retrieval and classification in a tiniest time as compared to existing algorithms concept and approaches. A concern with the data retrieval and classification, our approach returns efficient feedback with respect to the existing approach. See in Table 2.

As the complex checking of the classification and analysis of dataset, we are representing a confusion matrix on the data structure. Confusion matrix suggests a comparative analysis of data classification and also relative index with all columns and rows. See in Table 3.

For the proof of accuracy and analysis of the training and test set, we have applied the concept of precision and recall with the help of cross validation classification. The complete dataset is divided randomly in subsamples. These subsamples are equally divided. Out of these subsamples, one is used as a test set and remaining are as training set. See in Table 4 and Fig. 3.

Table 3 AnaData: confusion matrix

Dataset type	Structure (%)	Unstructured (%)	Semi-structured (%)	Quasi-structured (%)	Quasi-unstructured (%)
Structure	**95**	01	01	1	2
UnStructured	1	**96**	1	1	1
Semi-structured	01	2	**94**	1	2
Quasi-Structured	2	1	1	**94**	2
Quasi-Unstructured	1	1	1	1	**96**



(a) AnaData: Accuracy and Ability with RFT **(b)** AnaData: Accuracy and Ability with SVM

Fig. 3 AnaData: accuracy and ability

5 Conclusion and Future Work

The proposed approach 'AnaData' is a more efficient framework for data classification and data analytics in big data domain. We have tested approximately 2.2 GB of robust dataset and find efficient output with respect to this dataset is approximately 95%. Our approach overcomes the disadvantages of RFT and SVM and checks the ability with accuracy of the proposed framework 'AnaData' with statistical measurement platforms like cross validation measurement, precision, recall, comparative analysis and confusion matrix. Our approach provides fast retrieval of data with higher data analysis and classification in a tiniest time as compared to existing algorithms concept and approaches. As efficient with the data classification and analytics, our approach secures fast processing of data analytics and classification when compared to many proposed researches. As a concern with confusion matrix, it is a comparative analysis of data structure and robustness of our approach. As for the accuracy of the approach, it gives the result as 95%. As for the future work, we will work on large number of samples with many other domains.

Table 4 AnaData: accuracy and ability

	RFT			SVM			AnaData		
	Precision	Recall	H-mean	Precision	Recall	H-mean	Precision	Recall	H-mean
Correctly classified	0.82	0.83	0.826	0.84	0.86	0.852	0.88	0.89	0.885
Incorrectly classified	0.80	0.81	0.804	0.81	0.83	0.823	0.82	0.86	0.854

References

1. Google Cloud and Big Data. https://cloud.google.com/bigquery/ (2017). Accessed 20 Oct 2017
2. Digital Innovation Mobile Big Data. www.digitalinnovationgazette.com/mobile/big/data/ (2017). Accessed 17 Oct 2017
3. Venturebeat Big Data Analytics. www.venturebeat.com/2015/01/22/big-data-and-mobile-analytics-ready-to-rule-2015/ (2017). Accessed 15 Oct 2017
4. Knowledge Hut Types Of Big Data. https://www.knowledgehut.com/blog/bigdata-hadoop/types-of-big-data (2017). Accessed 12 Sept 2017
5. Impact Radius The seven Vs. https://www.impactradius.com/blog/7-vs-big-data/ (2017). Accessed 12 Nov 2017
6. Tang, J., Liu, J., Zhang, M., Mei, Q.: Visualizing large-scale and high-dimensional data. In: Proceedings of the 25th International Conference on WWW, pp. 287–297 (2016)
7. Yang, C., Huang, Q., Li, Z., Liu, K., Hu, F.: Big Data and cloud computing: innovation opportunities and challenges. International Journal of Digital Earth Published by Informa UK Limited, Trading as Taylor Francis (2016)
8. Shrivastava, A., Verma, V.K., Shankar, V.G.: XTrap: trapping client and server side XSS vulnerability. In: 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, pp. 394–398 (2016). https://doi.org/10.1109/pdgc.2016.7913227
9. Shankar, V.G., Somani, G., Gaur, M.S., Laxmi, V., Conti, M.: AndroTaint: an efficient android malware detection framework using dynamic taint analysis. In: 2017 ISEA Asia Security and Privacy (ISEASP), Surat, pp. 1–13 (2017). https://doi.org/10.1109/iseasp.2017.7976989
10. Brinker, K.: Incorporating diversity in active learning with support vector machines In: Proceedings of the 20th International Conference on Machine Learning, pp. 59–66. ACM, Washington, USA (2003)
11. Wager, S., Hastie, T., Efron, B.: Confidence intervals for random forests: the Jackknife and the Infinitesimal Jackknife. J. Mach. Learn. Res. JMLR **15**(1), 162–551 (2014)
12. Scornet, E., Biau, G., Vert, J.: Consistency of random forests. Ann. Stat. **43**(4), 1716–1741 (2015). https://doi.org/10.1214/15-aos1321
13. Kune, R., Konugurthi, P.K., Agarwal, A., Chillarige, R.R., Buyya, R.: The anatomy of big data computing. Softw. Pract. Exper. **46**, 79–105 (2016)
14. Shankar, A., Gauri, V., Somani, G.: Anti-Hijack: runtime detection of malware initiated hijacking in android. Procedia Comput. Sci. **78**, 587–594 (2016). https://doi.org/10.1016/j.procs.2016.02.105
15. Breiman, L.: Random forests. In: 2001 Random Forests Machine Learning, vol. 45(1), pp. 5–32 (2001)
16. Bernard, S., Heutte, L., Adam, S.: Forest-RK: a new random forest induction method. In: Proceedings of 4th International Conference on Intelligent Computing: Advanced Intelligent Computing Theories and Applications with Aspects of Artificial Intelligence. Springer (2008)
17. K Dnuggets Social Media Data. https://www.kdnuggets.com/2014/08/interesting-social-media-datasets.html (2017). Accessed 04 Oct 2017
18. UCI Transactional. http://archive.ics.uci.edu/ml/datasets/online+retail (2017). Accessed 28 Oct 2017
19. IMDB Movies Dataset. https://www.kaggle.com/orgesleka/imdbmovies (2017). Accessed 21 Oct 2017
20. Google Public Dataset. https://cloud.google.com/public-datasets/ (2017). Accessed 19 Oct 2017
21. Rtwilson Geographical Dataset. https://freegisdata.rtwilson.com/ (2017). Accessed 13 Oct 2017
22. MIT Genetic Dataset. https://www.ll.mit.edu//ideval/data/ (2017). Accessed 17 Oct 2017
23. Shankar, V.G., Jangid, M., Devi, B., Kabra, S.: Mobile Big Data: Malware and Its Analysis. In Proceedings of First International Conference on Smart System, Innovations and Computing. Smart Innovation, Systems and Technologies, vol. 79, pp. 831–842, Springer, Singapore, (2018) https://doi.org/10.1007/978-981-10-5828-8_79

A Decision Support System Using Analytical Hierarchy Process for Student-Teacher-Industry Expectation Perspective

S. S. Pawar and R. R. Rathod

Abstract Communication gap between expectations or requirements of student, teacher and industry is major issue for every engineering institute as well as for nation. It is necessary to make engineering or professional students skilled and employable for industries. Therefore, there is a need of proper understanding between student, teacher and industry with respect to various skills and making them aware of various engineering, professional and management practices and methodologies. The National Institutional Ranking Framework (NIRF) of Government of India (GoI) provides ranks for institutes based on various parameters however the proposed study focuses on common perspectives of student-teacher and industry for better employability, understandings and interactions. One of the parameter 'Graduation Outcomes' of NIRF has been used in present study. Analytical Hierarchy Process (AHP) has been applied to identify common perspective on expectations (POE) of Student, Teacher and Industry (S-T-I) for bridging the perspective gap using S-T-I survey data. The obtained result shows that there is a gap in expectations for few identified criterias among S-T-I. However these gaps can be minimized by increasing communication among S-T-I's.

Keywords Analytical hierarchy process · Criterias

1 Introduction

Industries have initiated various activities to establish relationship with institutes to accomplish different knowledge-, research- or innovation-based goals. Hence, it is desired to make engineering or professional students skilful and employable for industries. The proper communication between students, teachers and industries can

S. S. Pawar · R. R. Rathod (✉)
Department of Information Technology, Walchand College of Engineering, Sangli, India
e-mail: rrr.wce@gmail.com

S. S. Pawar
e-mail: shubhadapawar2117@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_54

expose students to newer technologies and make them aware of various engineering, professional, management practices. This can be achieved only by bridging the expectation gap between student, teacher and industry.

Ministry of Human Resource Development, Government of India has prepared National Institutional Ranking Framework (NIRF) since 2015 to rank institutions across the country with following parameters. Broadly it includes teaching learning and resources, research and professional practices, outcomes from graduates, outreach and inclusivity and perceptions of employers and research investors, academicians and public [1].

The present case study focuses on minimizing the communication gap of expectations between the students, teachers and industries using AHP.

The AHP is designed to solve complex problems involving multiple criteria used for measuring the qualitative and quantitative factors [2] in decision making. It helps and facilitates decision making based on judgments, feelings, memories and other forces that may influence decision making at multilevel hierarchy structures. In AHP, weights for each criterion are taken as input by the experts or user. There are three principle guides for problem solving using AHP—decomposition, comparative judgments, synthesis of priorities.

2 Related Work

A framework of AHP method is based on ordinal scale [2]. Saaty introduced Analytic hierarchy process (AHP) for decision-making [3] which is multi-criteria decision-making (MCDM) method created to meet the immense difficulties of choice circumstances that were brought by multiple (or even) criterias.

Mann describes use of AHP in engineering application [4]. AHP can be also used for various application areas such as industrial engineering applications, that in integrated manufacturing, investment decisions, flexible manufacturing systems, layout design, upgrade the computer system and also in other engineering problems.

AHP can be used as a potential decision making method for project management [5]. Proportion judgment scale is used for AHP method [6] and introduces a new method based on the proportion scale for construction comparison matrix in AHP. The study provides an overview of applications [7] where AHP is used such as in forecasting, medicine and related fields, personal, manufacturing, industry, social, education. AHP use is rising in developing countries. Benitez [8] proposed a method which enables balancing consistency and expert judgment to overcome weakness such as a lack of consistency within AHP. AHP can be used to choose the alternative that achieves the best compromise solution in irrigated areas [9] between a set of socio-economic and environmental attributes and also for socially sustainable supplier selection [10] through social parameters such as equity, health and safety, wages, education, philanthropy, human rights, Child and bonded, labour housing, ethics.

In various real cases, AHP is used to evaluate how the criterias are being defined and measured [11]. On the basis of problem type the selection of criteria arise from the organization's expertise in the ranking of alternatives or indicators. Prakash used AHP method for selection of third party reverse logistic (RL) partner [12] considering evaluation criterias such as capacity criteria, financial ability, IT system, service quality, RL activities, geographical location, partner image and experience and also for the selection of strategic alliance partner [13] with various evaluation criterias like equipment's, marketing and service, integration & network, finance, IT systems and logistics along with their sub-criteria.

Other than this, AHP can also be used in different areas such as for evaluating travel mode competitiveness in long-distance travel [14], to build a strategic framework for technology road mapping [15], Local Tourism Transportation Mode [16], E-commerce websites [17] etc.

A single most preferred option and ranking of options can be identified by using MCDM techniques. In complex decision problems, AHP method is a robust and flexible MCDM technique which divides a complicated system into a hierarchical process of elements, which contains objectives, evaluation criteria and alternatives. The simple method AHP able to give results similar to the ones obtained with the complex outranking method PROMETHEE and TOPSIS [18, 19]. The main advantage of the AHP that it is simple and fast understandable methods for people who are not aware of the MCDM techniques.

NIRF outlines a methodology to rank institutions across the India. The parameters broadly cover teaching, learning and resources, research and professional practices, graduation outcomes, outreach and inclusivity and perception [1].

In India, research has been carried out for finding important critical factors for effective Industry-Institute Interaction [20]. The Study identifies different critical factors and ranks are given of these factors by applying AHP method [21, 22]. This can be useful for policy makers for developing the plan of valuable guidelines to promote effective Industry-Institute Interaction in India.

3 Introduction to AHP

General AHP method (Fig. 1) is carried out in following six steps.

Step 1: Goal definition and identification of the set of criterias that are associated with the identified goal.

Step 2: The relative weights of criteria, sub-criteria, alternative solutions are determined by the comparison matrix. The criterias are compared in pairs and the pair-wise comparison matrix is obtained. The weights given to the criterias as per Saaty's ratio scale for pair-wise comparison matrix [2] as given in Table 1.

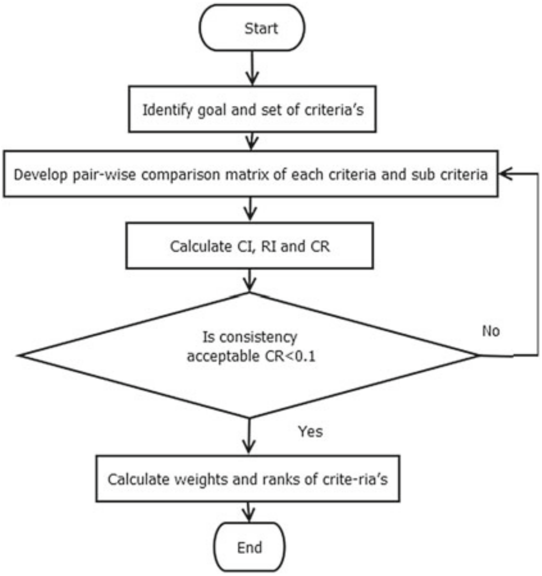


Fig. 1 Flowchart for general AHP method

Table 1 Saaty's ratio scale for pair-wise comparison of importance of weights

Intensity of importance	Definition
1	Equal importance
3	Moderate importance
5	Essential or strong importance
7	Very strong importance
9	Absolute importance
2, 4, 6, 8	Intermediate values between the two adjacent judgments

Step 3: Pair-wise comparison matrix at each level is constructed and normalizes the comparison matrix for calculating a consistent Eigen vector, consistency indexes (CI), random index (RI) and consistency ratio (CR).

i. For normalizing the matrix divide the weights of comparison matrix by the total value of their respective column. The average of normalized values on each row is equal to relative weight.
ii. Determine Eigen value, i.e., λ_max by calculating average of multiplication between matrix of relative weights and average weights of criteria.
iii. Compute values of random index (RI) from Table 2.
iv. Consistency Index (CI) is calculated using Eq. 1,

$$CI = \frac{\lambda \mathrm{max} - \mathrm{n}}{\mathrm{n} - 1}, \tag{1}$$

Table 2 Values of the random index (RI)

n	2	3	4	5	6	7	8	9	10
RI	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.51

where, λ_max is the maximum Eigenvalue of the judgment matrix and n is number of criteria.

v. Consistency ratio (CR) is calculated using Eq. 2,

$$CR = \frac{\text{CI}}{\text{RI}} \tag{2}$$

It is used to find out consistency of matrix CR value should be below 0.1.

Step 4: Based on the criteria and sub-criterias the alternative solutions are evaluated and ranked as the same via comparison matrix.

Step 5: On the basis of the weights ranks are given to the criteria. The consistency ratio (CR) of the alternative is determined and the alternative solution for the problem is identified.

4 Methodology

Considering a case study of Student-Teacher-Industry (S-T-I) perspective on expectation the methodology is carried out in two phases: the first phase describes the survey data collection process and the second phase explains the use of AHP.

4.1 Survey Data Collection

The online survey has been collected and weighing the criterias is carried out after identification of criterias rank. The detailed processing has been elaborated as follows.

Identification of Criterias: The identified criteria should have to match with the expectations of students, teachers and industries and hence feedback is collected from industries (HR), teachers and students. On the basis of technical aspects criterias were identified and literature review has been also carried out to identify various other criterias. These identified criterias are confirmed by industry (HR) and also some additional important criterias has been added as per their suggestions. Table 3 shows 18 criterias which are used for collecting survey responses.

Survey data collection: Considering each one's expectations survey data is collected from this entire three for weighing the criterias at different institutes and industries.

Table 3 Criterias for survey responses with weights

Sr. No	Criteria	Student	Teacher	Industry
1	**Practical's**	**9**	**9**	**9**
2	**Communication Skill (Oral)**	**9**	**9**	**9**
3	**Technical Skill**	**9**	**8**	**9**
4	**Knowledge beyond Textbook**	**9**	**8**	**9**
5	**Latest Technology**	**9**	**8**	**9**
6	**Programming Skill**	**9**	**8**	**8**
7	**Problem Solving Approach**	**9**	**7**	**9**
8	**Placement**	**9**	**7**	**9**
9	**Aptitude Skill**	**8**	**7**	**9**
10	**Leadership qualities**	**8**	**9**	**7**
11	Extra-Curricular Activities	7	8	7
12	Sports	7	8	7
13	Email Writing skill	7	7	9
14	Package	9	7	7
15	Placement Eligibility Criteria	8	5	9
16	Knowledge of Foreign Language	7	7	7
17	Classroom Teaching (Lectures)	7	8	8
18	Willing to work for shifts	7	8	7

The survey consists of 19 effective questionnaires where individual have to give weights between 1 and 9 to each criterion as per their own expectations and requirements where 1 represents low importance for that criterion and 9 represents high. In 19th question order of criteria has to decide.

4.2 *Student-Teacher-Industry Perspective on Expectation and AHP*

Industries have various activities to establish relationship with institutes to accomplish different knowledge, research or innovation based goals. It has been observed that very less number of students got hired after graduation; hence it is desired to make engineering or professional students skilful and employable. To produce proficient graduates ready for the industry, it is necessary to know the requirements and expectations of the industries and to establish a proper communication between student, teacher and industry at a common platform. This will make a great impact on the engineering curriculum design and exposure of industrial standards to engineering students and subsequent employability of engineer graduates in industries

Table 4 Weight calculations of students survey responses for criteria ‘aptitude skill’

Weights	1	2	3	4	5	6	7	8	9	Max = 8
Student Responses	0	1	1	1	13	14	53	48	39	48
Total	0	2	3	4	65	84	371	384	351	384

across the globe. Employability may achieve thus bridging the gap between student, teacher and industry by considering the expectation from each of them is necessary.
Goal: To find out common perspective of students, teacher and industry on employability.
Data analysis: Initially 18 criterias were identified however there is need of minimization of criterias for further studies, which has higher importance are considered.

Weights of criterias are calculated for student, teacher, and industry with respect to the number of responses; identified maximum weight will be the final weight for that criteria Table 4 shows identified final weight of the criteria aptitude skill. Table 3 shows weights of all criterias with respect to number of responses. In Table 3 top ten criterias are identified after analysing highest top ten weights average. Hence AHP method has been applied for these top ten criterias.

5 Results and Discussion

CR and ranks of top ten criterias calculated. For the results and following steps has been performed.

Step 1: Comparison matrices of top ten criterias are calculated with respect to student, teacher, and industry. Table 5 shows comparison matrix of only teacher.
Step 2: Normalization is applied to matrices and weights are calculated shown in Table 6. After calculation of CR, result says that CR of all three matrices is below 55%.
Step 3: Comparison matrix for student, teacher and industry is computed (Table 7) after normalization, weights were calculated.
Step 4: Calculate total value for finding rank by multiplying each weight of criteria related to student, teacher, and industry with the weights from comparison matrix (Table 8). Global ranking will get after performing summation of weights in row (Table 8). Practical’s have 1st rank.
Step 5: For finding communication gap, pair-wise comparison matrices between student, teacher, and industry are computed. Table 9 shows comparison matrix for Student-Teacher only.
Step 6: Follow the same steps for obtaining ranks of criterias for comparison matrices of student-industry, student-teacher and teacher-industry.

Table 5 Comparison matrix of criterias for teacher

Criteria	1	2	3	4	5	6	7	8	9	10
1	1	9	9	9	9	9	9	9	9	9
2	0.11	1	9	9	9	9	9	9	9	9
3	0.11	0.11	1	8	8	8	8	8	8	0.11
4	0.11	0.11	0.13	1	8	8	8	8	8	0.11
5	0.11	0.11	0.13	0.13	1	8	8	8	8	0.11
6	0.11	0.11	0.13	0.13	0.13	1	8	8	8	0.11
7	0.11	0.11	0.13	0.13	0.13	0.13	1	7	7	0.11
8	0.11	0.11	0.13	0.13	0.13	0.13	0.14	1	7	0.11
9	0.11	0.11	0.13	0.13	0.13	0.13	0.14	0.14	1	9
10	0.11	0.11	9	9	9	9	9	9	9	1

Table 6 Normalization of matrices

Criteria	1	2	3	4	5	6	7	8	9	10
Student	30.66	18.87	14.01	10.77	8.33	6.37	4.73	3.32	1.97	0.98
Teacher	31.19	19.4	14.54	10.17	7.96	6.14	4.61	3	1.96	1.02
Industry	30.66	18.87	14.01	10.77	8.33	6.37	4.73	3.32	1.97	0.98

Table 7 Comparison matrix for S-T-I

	Student	Teacher	Industry	Weights
Student	1.00	8.01	8.01	0.536
Teacher	1.00	1.00	3.57	0.214
Industry	1.00	5.45	1.00	0.25

Table 8 Ranks of ten criterias with respect to the student, teacher and industry

Sr. No.	Criteria	Student	Teacher	Industry	Average	Rank for S-T-I
1	Practical’s	0.164	0.067	0.076	0.103	1
2	Communication skill (Oral)	0.101	0.042	0.047	0.064	2
3	Technical skill	0.075	0.022	0.035	0.044	3
4	Knowledge beyond textbook	0.058	0.017	0.027	0.034	4
5	Latest technology	0.045	0.013	0.021	0.027	5
6	Programming skill	0.034	0.01	0.005	0.017	6
7	Problem solving approach	0.025	0.006	0.016	0.016	7
8	Placement	0.018	0.004	0.012	0.012	9
9	Aptitude skill	0.011	0.002	0.008	0.008	10
10	Leadership qualities	0.005	0.031	0.002	0.013	8

Step 7: Compare ranks of student-teacher, student-industry and industry-teacher with ranks obtained on common perception of student, teacher and industry (Table 10).

Step 8: It has been noticed the difference between the ranks where ‘+’ sign defines need of interaction to improvise in respective criterias whereas ‘−’ suggest to decrease the importance of criterias with respect to other’s perspective for common understandings. The ranking suggests that the criterias which needed communication and a few criterias where all three have same or common perspectives as shown in Table 10.

Student, teacher and industry are commonly accepting communication skill, practical’s, technical skill, knowledge beyond textbook, latest technology with equal importance and huge communication gap for other criterias has been identified.

Table 9 Comparison matrix for student-teacher

Student-teacher	1	2	3	4	5	6	7	8	9	10	Weights	Rank
1	1	9	9	9	9	9	9	9	9	9	31.51	1
2	0.11	1	9	9	9	9	9	9	9	9	19.69	2
3	0.11	0.11	1	8	8	8	8	8	8	8	13.34	3
4	0.11	0.11	0.13	1	8	8	8	8	8	8	10.38	4
5	0.11	0.11	0.13	0.13	1	8	8	8	8	8	8.1	5
6	0.11	0.11	0.13	0.13	0.13	1	8	8	8	8	6.24	6
7	0.11	0.11	0.13	0.13	0.13	0.13	1	7	7	0.13	1.04	8
8	0.11	0.11	0.13	0.13	0.13	0.13	0.14	1	7	0.13	4.68	9
9	0.11	0.11	0.13	0.13	0.13	0.13	0.14	0.14	1	0.13	3.04	10
10	0.11	0.11	0.13	0.13	0.13	0.13	8	8	8	1	1.99	7

Table 10 Communication gap between S-T-I

Sr. No.	Criteria	Interaction need		
		Student-teacher	Student-industry	Teacher-industry
1	Practical's	No	No	No
2	Communication skill (Oral)	No	No	No
3	Technical skill	No	No	No
4	Knowledge beyond Textbook	No	No	No
5	Latest technology	No	No	No
6	Programming Skill	No	Yes (−2)	No
7	Problem solving approach	Yes (−1)	Yes (+1)	No
8	Placement	No	Yes (+2)	Yes (+1)
9	Aptitude skill	No	Yes (+1)	Yes (+1)
10	Leadership qualities	Yes (+1)	Yes (−2)	Yes (−2)

6 Conclusion and Future Scope

In present study, AHP has been employed to identify and suggests performance evaluation criterias which can be improved by increasing effective communication between student, teacher and industry for identified lags in criterias. After comparing the ranks of pairs with ranks according to student, teacher and industry; results indicate that some have equal importance among all however a huge diverse perspective on criterias—problem solving approach, Placement, Aptitude Skill and Leadership qualities. Other criterias also need interaction for common perspective and hence needed to explore further in detail.

Serial AHP analysis is time consuming hence to achieve speed-up; AHP operations can be parallelized with more survey participants for better data analysis and result accuracy.

References

1. www.nirfindia.org. Accessed 01 Jan 2018
2. Wakabayashi, T., Itoh, K., Mitamura, T., Ohuchi, A.: A framework of an analytic hierarchy process method based on\nordinal scale. In: Proceedings of IEEE 5th International Fuzzy Systems. pp. 355–360 (1996)
3. Saaty, R.W.: The analytic hierarchy process-what it is and how it is used. Math. Model **9**, 161–176 (1987). https://doi.org/10.1016/0270-0255(87)90473-8
4. Triantaphyllou, E., Mann, S.H.: Using the Analytic Hierarchy Process for Decision Making in Engineering Applications: Some Challenges. Int. J. Ind. Eng. Theory Appl. Pract. **2**, 35–44 (1995)
5. Al-Harbi, K.M.A.-S.: Application of the AHP in project management. Int. J. Proj. Manag. **19**, 19–27 (2001). https://doi.org/10.1016/S0263-7863(99)00038-1

6. Decai, H.: the Proportion Scales in the AHP. J. Syst. Eng. Electron. **14**(3), 8–13 (2003)
7. Vaidya, O.S., Kumar, S.: Analytic hierarchy process: an overview of applications. Eur. J. Oper. Res. **169**, 1–29 (2006). https://doi.org/10.1016/j.ejor.2004.04.028
8. Benítez, J., Delgado-Galván, X., Gutiérrez, J.A., Izquierdo, J.: Balancing consistency and expert judgment in AHP. Math. Comput. Model. **54**, 1785–1790 (2011). https://doi.org/10.1016/j.mcm.2010.12.023
9. Gallego-Ayala, J.: Selecting irrigation water pricing alternatives using a multi-methodological approach. Math. Comput. Model. **55**, 861–883 (2012). https://doi.org/10.1016/j.mcm.2011.09.014
10. Mani, V., Agarwal, R., Sharma, V.: Supplier selection using social sustainability: AHP based approach in India. Int. Strateg. Manag. Rev. **2**, 98–112 (2014). https://doi.org/10.1016/j.ism.2014.10.00
11. Russo, R.D.F.S.M., Camanho, R.: Criteria in AHP: A systematic review of literature. Procedia Comput Sci **55**, 1123–1132 (2015). https://doi.org/10.1016/j.procs.2015.07.081
12. Prakash, C., Barua, M.K.: An analysis of integrated robust hybrid model for third-party reverse logistics partner selection under fuzzy environment. Resour. Conserv. Recycl. **108**, 63–81 (2016). https://doi.org/10.1016/j.resconrec.2015.12.011
13. Garg, C.P.: A robust hybrid decision model for evaluation and selection of the strategic alliance partner in the airline industry. J. Air Transp. Manag. **52**, 55–66 (2016). https://doi.org/10.1016/j.jairtraman.2015.12.009
14. Wang, Y., Yan, X., Zhou, Y., Li, X.: Using AHP for evaluating travel mode competitiveness in long-distance travel. In: ICTIS 2015—3rd International Conference on Transportation Information and Safety, Proceedings, pp. 213–218 (2015). https://doi.org/10.1109/ictis.2015.7232180
15. Gerdsri, N., Kocaoglu, D.F.: Applying the analytic hierarchy process (AHP) to build a strategic framework for technology roadmapping. Math. Comput. Model. **46**, 1071–1080 (2007). https://doi.org/10.1016/j.mcm.2007.03.015
16. Jin, H., Yang, X.: Efficient organization study of local tourism transportation mode based on analytic hierarchy process. In: 2014 17th International Conference on Intelligent Transportation Systems ITSC, pp. 1555–1560 (2014) . https://doi.org/10.1109/itsc.2014.6957914
17. Anand, O.: Ecommerce AO, In P E-Commerce Website: a Hybrid Mcdm, Eighth International Conference on Contemporary Computing (IC3), Noida, 07 Dec 2015, pp. 279–284 (2015)
18. Tscheikner-Gratl, F., Egger, P., Rauch, W., Kleidorfer, M.: Comparison of multi-criteria decision support methods for integrated rehabilitation prioritization. Water (Switzerland) **9** (2017). https://doi.org/10.3390/w9020068
19. Kumar, S., Luthra, S., Haleem, A.: Critical factors important for effective industry-institute interactions (Iii): an Indian perspective. Int J Anal Hierarchy Process **8** (2016). https://doi.org/10.13033/ijahp.v8i1.319
20. Dodgson, J.S.: Multi-criteria analysis: a manual. Department for Communities and Local Government, London (2009)
21. Baharin, M., Ismail, W.R., Ahmad, R.R., Majid, N.: Factors affecting students' academic performance using analytic hierarchy process (AHP) In: 2015 International Conference on Research and Education in Mathematics (ICREM7), pp. 169–173. IEEE (2015)
22. Sahroni, T.R., Ariff, H.: Design of analytical hierarchy process (AHP) for teaching and learning. In: 2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS). IEEE (2016)

English Language Adoptability in Engineering Graduates: A Case Study

Sushama Deshpande, Amit Shesh and Brijesh Iyer

Abstract Unanimously accepted as Lingua Franca, English language has attained the status of global language with an important role in our daily life. Non-technical skills like problem solving, interpersonal skills, critical and independent thinking, positive attitude, active listening, a trait of enthusiasm, etc. are very important for Engineers. The prominent among all these non-technical skills is English communication skills. Engineers may become obsolete if they do not possess good communication skills in English. This paper reports the empirical study of engineering graduates regarding their adoptability towards English language in practice. The analysis was carried out with the help of a comprehensive questionnaire. It is observed that the problem of English communication can be overcome easily with a systematic approach of teaching-learning process for engineering graduates.

Keywords Communication skills · Language learning skills

1 Introduction

Engineers form the backbone of the industry work force in this rapidly changing competitive world. They should have the skills and talent that make them employable once they enter the job market. Many engineers feel well prepared for their first job after completing their degree course. But they may be wrong. Acquisition of mere technical skills may make an engineer stand nowhere if his skills are not supported by non-technical skills. The corporate world demands that its would be

S. Deshpande (✉) · A. Shesh
Department of English, Dr. BATU Lonere, Lonere, India
e-mail: sushamabindu@gmail.com

A. Shesh
e-mail: apshesh@dbatu.ac.in

B. Iyer
Department of E & TC Engineering, Dr. BATU Lonere, Lonere, India
e-mail: brijeshiyer@dbatu.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_55

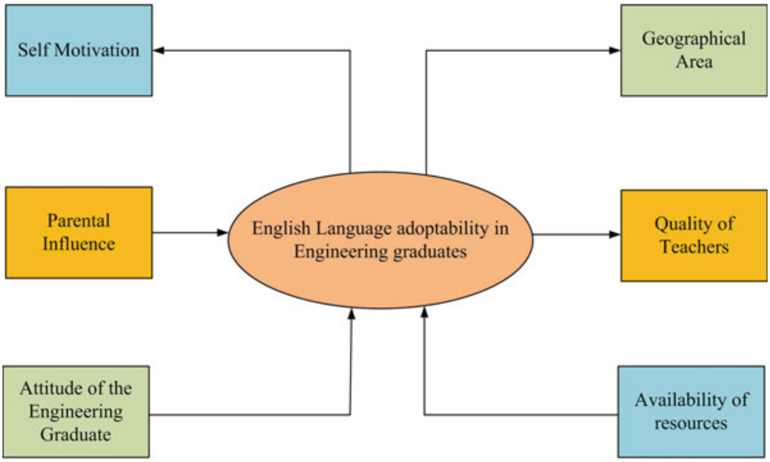


Fig. 1 Adoptability of English in Engineering Education

employees should have effective communications skills and soft skills. It is very important for the engineers to possess required communication skills and propose their thoughts effectively. Along with oral communication written communication is also very important for engineers. Today, engineers who can write clearly, concisely and comprehensively are in demand. Oral presentation skills are also an increasingly integral part of engineering profession. Good presentation skills and delivery style will add few more marks to an engineer's credit during his evaluation for a job.

It is found that English proficiency of some Engineering graduates is at par while some others are very weak in using English language. It means there are many factors which affect the English proficiency of Engineers. Through this study, an attempt is made to analyse these factors.

The rest of the paper is organised as: Sect. 2 reports literature review and theoretical background for the proposed study whereas the statement of the problem and experiment methodology is discussed in Sect. 3. The observations and allied conclusions are discussed in Sect. 4. The paper concludes in Sect. 5 with the brief commentary on the future scope for this work.

2 The Literature Review

2.1 Functions of Language

The act of expressing our feelings and emotions by the way of speech makes man different from other animals and superior in the universe. The act of verbal communication is the greatest power that the human beings have had. The contemporary

American philosopher David Abram wrote, "Only if words are felt, bodily presences, like echoes or waterfalls, can we understand the power of spoken language to influence, alter, and transform the perceptual world" [1]. The major function of any language is to express and communicate. Language is used to ask questions, reply, greet each other, describe people, things, ideas and our observations, and express our thoughts and emotions. When someone speaks, the speaker wants to convey some message. If his message is passed to the listener successfully, the communication process is successful. Without words or linguistic expressions, this communication process may not be very successful though non-verbal communication plays very important role in communication.

2.2 Importance of English Language

Language is an important tool of communication. In the present era, English language has the status of global language which is widely used language in the world. It is a global language which is used as first, second or third language in different countries. Due to its adaptable nature, English language got enriched with addition of many words from different languages. The basic spread of the language was due to British colonialism and today it is also regarded as the language of technical education. It is the need of the hour for everyone to learn the language. One who is interested in higher or technical education cannot run away from the language. English language is used as the medium of instruction in all the streams as well it is used for day-to-day communication. Apart from communicating with others, Engineering graduates have to use language for reading technical books and acquiring subject knowledge, for written and oral presentations, conducting research and publish their findings. Along with the academic proficiency the engineers have to possess high level of proficiency in communication. Due to poor communication skills the engineers would be deprived of the job opportunities.

2.3 Basic Language Learning Skills—Macro Skills—LSRW

The basic function of a language is to communicate. There are four language learning skills most important for oral and written communication. They are Receptive Skills like Listening and Reading and Productive Skills like Speaking and Writing. Students can understand and process the language and respond to the messages by using these skills. If a student wants to justify his professional supremacy and compete with his counterparts, he must have a mastery over these macro language skills.

All children are born with the ability to learn language from their childhood. Before a child learns his mother tongue formally in school, he starts learning it by constantly absorbing and acquiring sounds from his surroundings. The child also imitates sounds and tries to repeat them. After getting this phonemic awareness, child learns words, phrases, sentence patterns and grammatical correctness and acceptability of sentences.

When it comes to learning a foreign language (English, in case of Indians), students face problems because they do not learn it by such natural process as they learn their mother tongue. They take the formal education of the language after they learn their mother tongue. They find it difficult to learn new cognitive and sound patterns as their brain is used to the patterns of their mother tongue. The process becomes more tiresome when students ignore the language in their school days and later find it indispensable for their career.

Many efforts were reported in the literature to address the issue of English language for the engineering graduates. In [2], criteria for the Engineering accreditation in the United States of America were reported. It discussed the importance of role of communication skills and employability in engineering graduates. The importance of communication ability for engineering apprentice was discussed in [3]. *Keane and Gibson* had discussed the communication trends in various engineering firms and its implications on engineering graduates [4]. *Yurtseven H* reported the role of English communication in recruitment and retention of job [5]. Role students' skills and experiences in quality engineering education were reported by *El-Raghy* in 1999 [6]. In this study he predicted that the communication (written and oral) ability in English will be the key in the coming decade for the engineering graduates. However, none of these efforts studied the cause–effect relationship for the adoptability of English language and allied communication ability in engineering graduates. The present work tries to carry out the cause–effect relationship study for the problem of English language in engineering graduates with the help of a well-defined questionnaire.

3 Statement of the Problem

The English language proficiency of students who enter engineering colleges may vary considerably. Hence, it is necessary to evaluate their capability to manage learning and using English in their academic and professional life. In the absence of adequate competency in English language skills, a majority of the learners face problems in mastering the technical terminologies of science and engineering textbooks written in English. Therefore, in order to make them comfortable while learning various engineering courses to keep pace with the latest trends in the field of science and technology, they require an adequate level of reading and comprehension skills in English.

It is seen, in fact, that many students are not able to understand and use English language up to the mark. In the same campus we find students with different levels of intelligibility. It is so because they come from varied background. Their background differs academically, socially, geographically and financially. These factors affect the students' ability to learn and apply English language to a great extent. Once the factors affecting the students' language competency levels are evaluated, it will become easier to devise teaching methodologies appropriate for different students and make the teaching-learning process more effective.

Hence, the factors affecting the English proficiency of engineering graduates are analysed and studied in this work.

3.1 Significance of the Study

1. In the age of globalisation, engineering students must enhance their English communication skills and soft skills in order to have more job opportunities in the world with tough competition in the job market. Apart from just possessing the subject knowledge, students should be able to deliver the same efficiently in front of the audience or readers which would increase the chances of employment.
2. Though English communication is important for the engineering students, it is found that not all the students are able to possess good communication skills.
3. Many internal and external factors affect the students' ability to learn English language and deliver knowledge they have acquired.
4. The present study aims at studying all these factors affecting the students' English communication skills so that the study can be further carried out to devise appropriate teaching methodology to improve English proficiency of engineering graduates.

3.2 Methodology

Students from different backgrounds register with the university to pursue higher education. Everyone's ability to use English language differs drastically. It is found that language acquisition pattern of each student is also different and affected by different factors. To analyze these factors, a questionnaire-based survey among the undergraduate and postgraduate students of engineering from Dr. BATU, Lonere was undertaken. Sample data of 150 students was collected for the project.

3.3 Questionnaires

Total 10 questionnaires were prepared through which both qualitative and quantitative data was collected. For preparing these questionnaires, Google forms were created and the data was collected through online mode. The objectives of these questionnaires were to find out students' attitude towards English language and the factors that have affected their learning of English. The questionnaires circulated among the respondents were in the form of multiple choice questions.

3.4 Analysis of the Data

All the students filled up the questionnaires. Initial hesitation of the students was overcome by explaining them the purpose of the study. Apart from gaining information about the factors affecting English proficiency, the questionnaire focused the need of English language for engineers and its awareness among the students. The study made use of 10 questionnaires. Sample included around 150 students of under graduation and post-graduation from Dr. BATU, Lonere.

The intention behind the questionnaire-based survey is to obtain information from a large number of participants in engineering in order to understand their beliefs and capabilities. Necessity to ensure the validity of the data collected should be stressed, because the findings of the research are based on this data. The data collected for this research is validated for its reliability and completeness at the time of colleting the filled up questionnaires through online mode.

4 Results and Discussions

Language learning is a psychological process which is greatly affected by various social, geographical, topological, attitudinal and many other factors. These factors can be divided into Internal and External factors. Before actually taking efforts to learn English language, students should be aware of the need to learn English to become successful in their professional life.

1. **Internal Factors**: Internal factors are the problems and situations that the learner faces at a particular learning situation. Following Internal factors affect the students learning of English language:

   (a) **Age**—Children acquire a new language easily while grownups take a lot of time to learn a new language.
   (b) **Competitive spirit**—The reality of 'Survival of the fittest' brings competitive spirit among the students which inspires them to enlarge the horizon of their skills and talent by learning new language.
   (c) **Attitude**—Some students develop positive attitude towards the use of English language and its need. Such students learn and use English language in their daily life whereas the students who develop negativity and unnecessary fear about English language develop a phobia for the language.
   (d) **Intelligence**—The basic intelligence level of students matters a lot while learning English language. Those who are intelligent, learn easily by concentrating but students who are not very intelligent, need to take a lot of efforts for language learning.

(e) **Fear/Anxiety**—Fear or inhibition is the psychological aspect that acts as a barrier during the process of communication. It is very common while learning or using a foreign language. Most of the students cannot overcome this problem due to lack of practice. As compared to the children, the adults are more conscious while using the language because they are worried about their image in public. Whereas, the children may take lot of liberty while using the language due to innocence and they may not care more about the judgement from the audience, the delivery of the content is more important for them.

(f) **Motivation**—Motivation is considered as an important factor in our day-to-day life. Motivation is an important factor in the teaching-learning process. It has been proved by the psychiatrists. The teacher has to boost the learners psychologically while teaching the language so that he would achieve his goal. The teacher can provide the learners many opportunities for learning and motivate them for learning the language. The teacher has to design the course in such a way that the learners would enjoy the teaching-learning process that will help them to enhance their linguistic abilities. The learners who enjoy language learning will do better than the peers.

(g) **Self-esteem**—It is one of the psychological factors that stimulates the learners, compels them enhance their linguistic abilities. Those who have self-esteem would try to use better English. Especially, in case of adults self-esteem plays an important role due to which the learners would be more cautious about the quality of the language they would use in front of their audience. Hence, it becomes an influential factor in the process of teaching and learning a language.

(h) **Openness to innovation and new methods**—An open mind can embrace changes and scientific innovations and new methods of leaning and innovations. Such students want to bring changes in themselves to adapt to the new environment whereas a rotten mind is always against discoveries and innovations and does not accept changes.

(i) **Personality Type—Extrovert & Introvert**—Students who are extrovert can express themselves and so learn English language easily whereas those who are introvert cannot get any platform to express themselves. It is difficult for them to express themselves in a new language. In order to learn the language, the learners should be exposed to different situations where they need to use and practice English language.

2. **External Factors**: There may be different external factors that affect the process of language learning. Following External factors affect the students' learning of English language.

   (a) **Geographical Area**—It can be found from the given data that students' who are born and brought up in urban area have a good English proficiency than the students having rural background. 78.4% of the participants felt that the geographical area plays an important role in adopting English.

   (b) **Vocabulary**—A good command on vocabulary gives a student an edge over other students. Students feel confident when they have a huge vocabulary in their store.

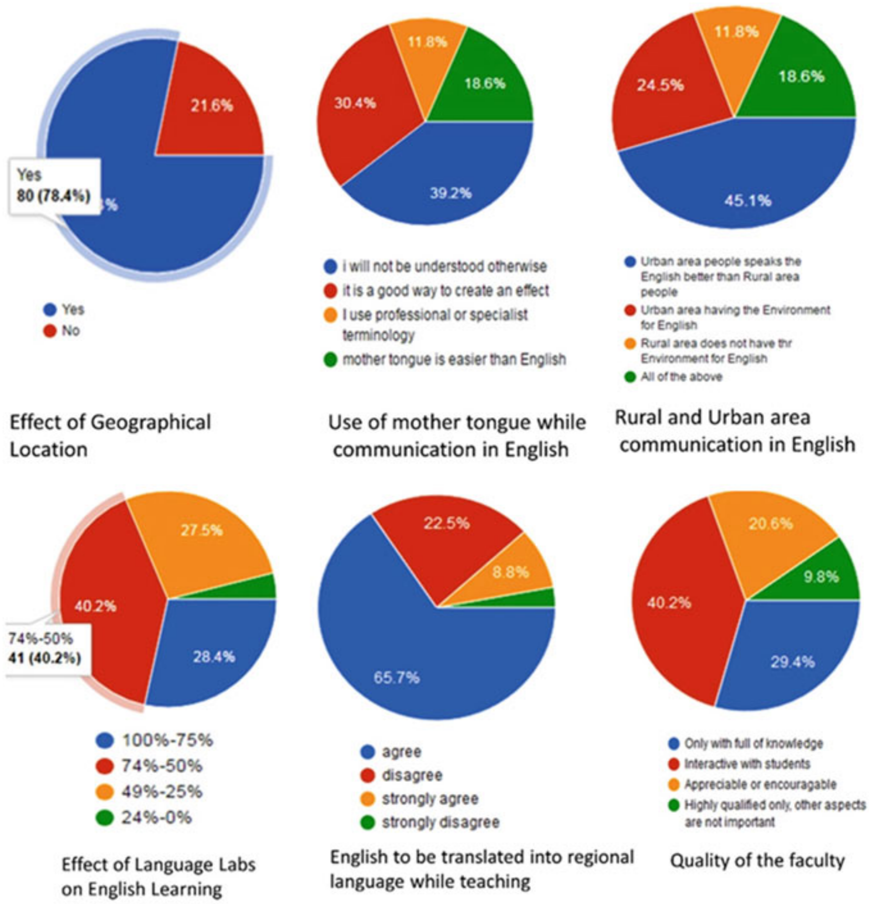


Fig. 2 Factors affecting and influencing the adoptability of English language in Engineering graduates

Friends—Friends are very important factor that has a continuous as well as a long-term effect. If friends themselves are interested in learning language, they are motivating or vice versa.

(d) **Parental influence**—If parents are educated and made aware of the necessity of English language in the career of their ward, they too take efforts or motivate the child to learn the language from the beginning.

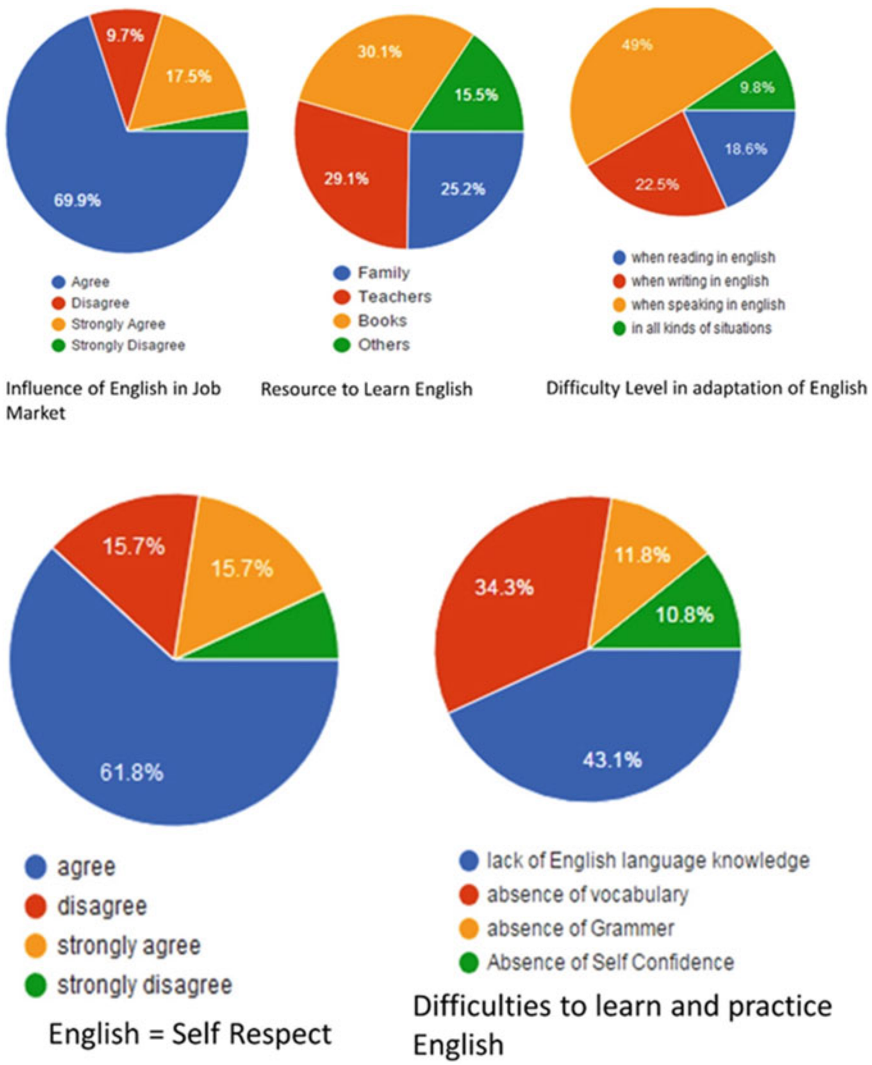


Fig. 2 (continued)

(e) **First language teacher** –First language teacher may arouse interest among students about English language by her innovative teaching methodology or create fear about the language.
(f) **Social Class**—Students who belong to that section of society where English is spoken regularly by all members, get a good environment to express themselves whereas the students who come from humble and illiterate background need to take individual efforts for learning.
(g) **Teaching and learning resources**—Students who get access to various resources like books, audio visual aids and internet can learn easily but students residing in rural area are deprived of all these resources which affect their language learning.
(h) **Career goal**—Students who aspire for a successful career are aware of the importance of English language and take efforts to learn it.
(i) **Curriculum**—Curriculum also plays a pivotal role in order to improve the communication skills of the students. If the curriculum is designed by giving importance to communication activities, students get opportunity to learn communicative language. Unfortunately, our education system which gives more importance to English language theoretically, gives less importance practically. Teachers give less exposure to the students to use language and students also want to learn the language without using it orally.
(j) **Job market**—Students learn English communication skills when they are very important from the point of view of their job. Technological jobs are very attractive but competitive as well. One who is completely prepared has chances to survive here (Fig. 2).

5 Conclusions

Based on the comprehensive questionnaire, the various factors influencing the adoption of English language in Engineering graduates were studied. It is found that geographical location, ability of the faculty to express and explain the concept in mother tongue, surrounding environment are the few key elements which can be worked out to improve the present-day scenario in Engineering graduates on English language front. Language laboratories should be incorporated in the curriculum along with classroom sessions, it would play a vital role in the teaching-learning process that is undertaken in the classroom. Laboratory sessions would include the interactive sessions based on role play, extempore, group discussions, presentation, practice, development and testing.

In future, the issue of English language adoptability will be analyzed by using Big-data analytics.

References

1. Abram, D.: The Spell of the Sensuous: Perception and Language in a More-Than-Human World. Vintage Books, A division of Random House, Inc., New York (1997). ISBN 978-0-679-77639-0
2. Baum, E.: Engineering accreditation in the United States of America—Criteria 2000. In: Proceedings of the 2nd Global Congress on Engineering Education, Wismar, Germany, pp. 17–20 (2000)
3. Jensen, H.P.: Strategic planning for the education process in the next century. Global J. Eng. Educ. **4**(1), 35–42 (2000)
4. Keane, A., Gibson, I.S.: Communication trends in engineering firms: implications for undergraduate engineering courses. Inter. J. Eng. Educ. **15**(2), 115–121 (1999)
5. Yurtseven, H.O.: How does the image of engineering affect student recruitment and retention? A perspective from the USA. In: Proceedins of the 4th UICEE Annual Conference on Engineering Education, Bangkok, Thailand, pp. 62–65 (2001)
6. El-Raghy, S.: Quality engineering education: student skills and experiences. Global J. Eng. Educ. **3**(1), 25–29 (1999)

Design and Development of E-Care Management System for Hospitals

Mrutyunjaya S. Yalawar, Basava S. Dhanne, Rakesh Ranjan and Telugu Satyanarayana

Abstract E-care management system is a system which provides support to doctor, patient, management and other stake holders with the required information to carry out their day-to-day routine work. It is a complete solution for recording and retrieving all possible transactions in a hospital. This system provides complete solution to patient for getting appointment from doctor, getting treatment. The main aim of this paper is to clarify the importance of linking hospitals electronically and sharing patient medical history records, so it becomes available and accessible for all authorized users throughout all the regional hospitals, which saves time by reducing queries about medical history and medical condition updates. In this paper we focused on administration modules, patient Module, doctor Module and billing Module. We developed an application named "Med Application". It provides required information to the operational management for planning and executing their operations. This intelligent application provides all the information to the management from their day-to-day routine work to their future planning to assist the patient in all the aspects.

Keywords E-care · Hospital information system · Medical history · Patient healthcare · E-linking

M. S. Yalawar (✉)
Department of CSE, CMR Engineering College, Hyderabad 501401, Telangana, India
e-mail: muttusy@gmail.com

B. S. Dhanne · R. Ranjan · T. Satyanarayana
Department of ECE, CMR Engineering College, Hyderabad 501401, Telangana, India
e-mail: basu.dhanne@gmail.com

R. Ranjan
e-mail: rakeshranjan385@gmail.com

T. Satyanarayana
e-mail: satyant234@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_56

1 Introduction

There is an increasing need to find efficient ways to ensure appropriate patient information transfer between hospitals. The information on each patient is not shared automatically among the hospitals and therefore patients have to copy their files and take them to other hospitals if they are asked to. This results in the issues where in the case of older patients; doctors sometimes find it difficult to know the exact details of the patient's case. Patients' medical records are scattered between hospitals and healthcare centers and cannot be shared between two or more different hospitals simultaneously. This can prevent physicians and specialists from being informed about the full medical condition of the patient which may result in a reduction of the quality of the provided care. Electronic linking will enhance this quality and enable patients to receive enhanced treatments.

According to the survey report of an Institute of Medicine (IOM), between 44,000 and 98,000 Americans die each year as a result of medical errors [1] that could have been prevented. Beyond the cost in human lives as a result of such errors, the loss of credibility and integrity of some medical centers as well as hospitals is also damaging [2, 3]. People tend to use internet increasingly in their lives; checking for internet connections periodically anywhere they may find themselves. Reports indicate that around 34% of total population in the India who own smart phones use it daily to access the internet. It enables them to follow up on their businesses periodically and catch up on any updates upon its ongoing synchronization.

The electronic linking of hospitals will bring about the possibility of accessing patients' medical records at all the hospitals of a particular region which will also allow the patient to have complete approved documentation of the medical reports to allow the effective following up of the patient at any one of the linked hospitals [1, 2, 4]. The e-linking among the hospitals could lead to effective diagnosis by doctors, as they will have full access to patients' medical history, increase the organizational communication and enhance integration between hospitals by facilitating the sharing of information and updates on patients.

The main purpose of this work is to clarify the importance of linking hospitals electronically and sharing patient medical history records, so it becomes available and accessible for all authorized users throughout all the hospitals, which saves time by reducing queries about medical history and medical condition updates [1, 5, 6]. The system will also make it easier to diagnose the patient's condition therefore negating. Overall, the use of health information technology via E-Care management system should lead to more efficient, safer, and higher quality healthcare; highlighting the importance of the use of this system in hospitals of a particular region [7, 8].

E-care management system provides the information about the complete hospital management starting from basic details to different facilities available in a hospital. Many other applications are created or made on the healthcare system and one application differs from other. During earlier days there were no computerized hospital management system, the records and observations are maintained using papers along with this billing and transaction also maintained unconventional methods that is leads to loss of data and this process is completely dependent on man power, and waste of time. For the fulfilment of the objective, we are developing Med application, which is an intelligence information and management system to assist and manages various assorted things of hospital like (supervisory details, medical and commercial) and it also assists to keep a record of daily transactions of hospital and as well as permitting staff to pay longer caring for patient.

2 Literature Survey

In current scenario people visit the doctor explains the symptoms he is suffering from and takes the required medicines upon the doctor's suggestion. If the same problem is faced by the patient and if the doctor is not available then he has to again visit another doctor for same problem he faced before and he has to again explain everything about the reports for that problem [9]. And again if patient faces the same problem he has to visit another doctor again and again. The drawbacks of the prevailing system:

- The hospital workers find it dull and time overwhelming once analyzing patient information. This ends up in bog down in medical reports.
- The hospital management presently uses health related files for storing patients and drug supplier's data. This technique of data storage is prone to security issues like no legal refinement and update of records.
- The workers sometimes waste plenty of time in retrieving information and paper wasting too.

In the sector of health wellness, lots of alternative user teams (Different department doctors, staff nurses, supervisor etc.) with style of framework and incompatible interest exist. There are nine modules in this hospital management system and each one of them are described below:

a. **Patient System**: This a kind of Hospital management System where in this Module all the Patients related information, history, details saved in this system. It includes operations like registration of the patients, viewing his all history, records related to the respective patients.
b. **Doctor System**: It is a kind of Hospital Management System which includes the details of the doctors, there registered patient's information along with the schedule and appointments of their respective patients.
c. **Drugs System**: In this module the overall information of the Drug system available along with the patient's information. It includes list of drugs used in the treatment for the specific hospital for a required patients.
d. **Administrative Rights System**: This Module includes the details of HR and administration related to the hospital management. It also updates the overall information of the patients and all other stocks in hospital.
e. **Online Appointment System**: In this Module it keeps the registration of new patients through online which avoids the physical visit for the hospital. This helps

in maintaining the patients, doctor appointments along with the status of patients in hospital.

f. **Invoice Billing System**: This is the module where once the appointment get confirmed, the invoice will be generated automatically with respect to the specific patient, which can able to show to all the payment, transaction details of the patients.
g. **Medical Services Management System**: This service system allows the inventory facilities given to hospital like various treatments dental, cardiac, bone treatment and many more. It manages all the facilities of the medical system.

As in the existing system the patients needs to visit the hospital for taking the appointment along with the concerned for the doctor [3, 5, 10, 11]. Due to the unavailability of the respective doctor, the patients need to go/visit for another doctor again and again due to lack of health records or the information of the patients. In this proposed system Patient register through online and get unique ID after completing registration process. Patient then consult doctor, medicines which were prescribed by doctor will be placed in patients' medical history, and patient can choose to buy medicine online and can have the benefit of medicine home delivery. The customer is provided with the view medical history option for the products available if he forgets that medicines.

3 Proposed System

In proposed system Patient register through online and get unique ID after completing registration process. Patient then consult doctor, medicines which were prescribed by doctor will be placed in patients' medical history, and patient can choose to buy medicine online and can have the benefit of medicine home delivery. The customer is provided with the view medical history option for the products available if he forgets that medicines. By developing this system we can save the valuable time of the customer and Med application is easy and user friendly for the patients as well as doctors. Figure 1 is depicting the data flow diagram of E-care management system.

3.1 Modules and Their Functionalities

There are four modules and each of them are defined below

- **User module**: This system is used for registration, accessing doctor's list, booking appointment, access patient's history, prescriptions, payment.
- **Admin module**: This Module used to update all the information regarding the patient's data, doctor information and Hospital management. It is a main module which manages all the activities of the employment of the Hospital.

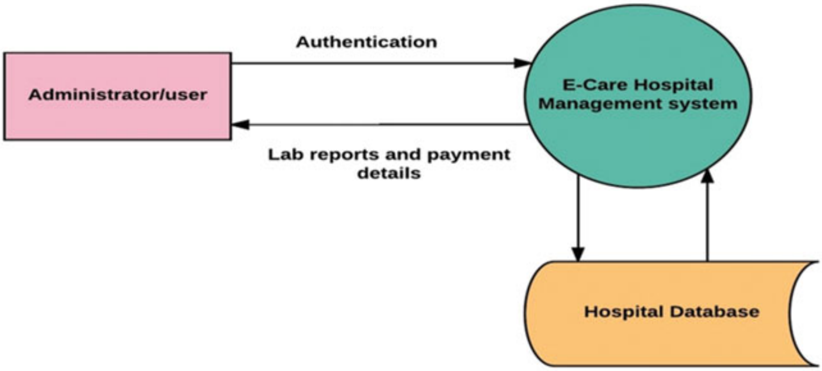


Fig. 1 Data flow diagram of E-care management system

- **Doctor module**: This Module consist of the details of all registered Doctors, even the complete information of their appointments and schedule carried out in hospital with respected patients details who have concerned already.
- **Patient module**: This module consists of the all the information of the registered patients, reports and detailed history. It allows the patients to know there all history during the treatment.

3.2 Functional Requirements

Our application includes the mainly Administration modules, Patient Module, Doctor Module and the Billing Module. The Administration Module usually operated on inserting, deleting, updating and viewing the patient's information in the application. The Patient Module will be operated on Patient Id, Name, Age, Address, and Phone Number. It can be again subcategorized into two sub modules as Inpatient module and the Outpatient Module. Usually in patient modules the details of the patients who are already admitted in the hospital as per the doctor advice. It involves Id of the patient, disease of the patient, name of the concerned doctor, Room number of the patient allocated. On the other side the Outpatient module includes the treatment already given for the patients, details of the disease. The Billing Modules includes the operations like doctor's charge, health card amount, room bill, medicine bill, total amount, No of days patients, Service charge, Operation theater, Nursing care, Lab bill. The Doctor Module includes the Doctor Name, specialization of the doctor, doctor Id, his background history, and patient's history details. The Application retrieves and sort information of the Medical records along with the output of the respective concerned patient's information.

Table 1 Test cases

Function	Expected results	Actual results	Remarks	Comments
Allow the patient to register	The patient should successfully register	The patient should register successfully	Success	If the patient does enter any fields then registration won't be successful
Allow the patient to login	The patient should be able to view the home page	The patient should be able to view the home page	Success	If the login and password are wrong the patient won't be able to view home page
Patient should be able to book appointment for doctor	The patient Should be able to select the slot	The patient Should be able to select the appointment timing	Success	By selecting u
Patient should be able to order and buy the medicine	If the account number is valid then the transaction should generated	If the account number is valid then the bill should be generated	Success	Bill will be generated successful

3.3 Non-functional Requirements

Usually our Application design allows location of the dropdown menus, options, and help. If the user tries to input for many times, input errors will be turned into bright blue with the validated message box. More than three attempts by the doctor, patient during the login and failure, it will arise a red flag alert to the administrator.

4 Results and Discussion

E-care management system provides the information about the complete hospital management starting from basic details to different facilities available in a hospital. In this work we focused on administration modules, patient Module, doctor Module and billing Module. We developed an application named "Med Application". It provides required information to the operational management for planning and executing their operations. We successfully developed the application and tested it. We tested "Med Application" for some cases. Table 1 is depicting the test cases.

We found some desired results. Figures 2 and 3 is showing different obtained output.

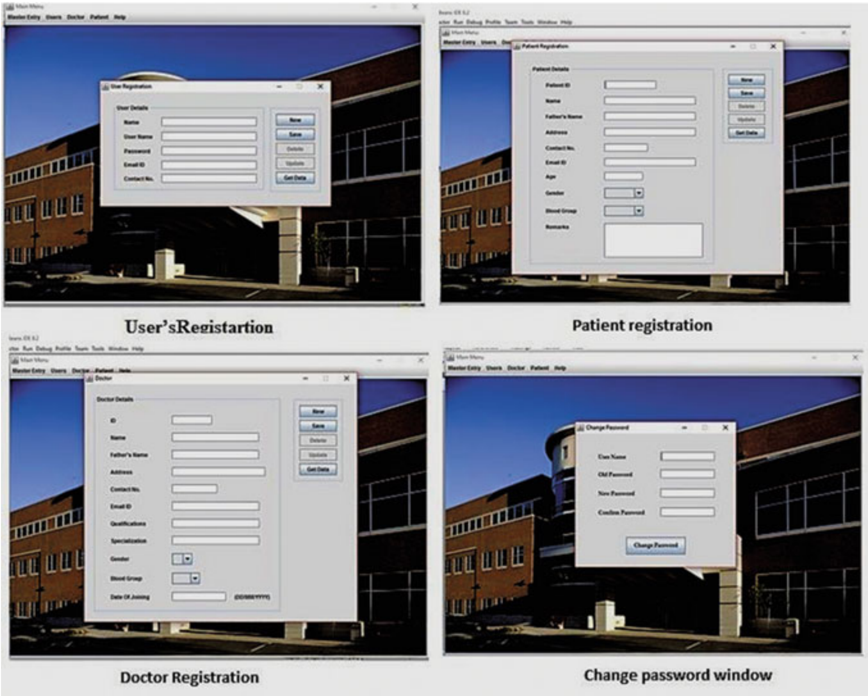


Fig. 2 Different users registration and change password window

5 Conclusion

The E-care management system is used to enhance and support the decision-making process that relates to patients' health. E-care system aimed to promote the healthcare sector in the community to accompany the challenges and the development around us. The designed "Med application" is a web based application. By using this application the patient can effectively register his profile and maintain his profile and also get medicine home delivery. This application offers the patient very flexible way registrations and using this system the patient can purchase medicines from his/her place which saves their time. Hopefully by using this system the community and healthcare sector will be served by making patients more aware about their health conditions. To develop this software application more efficiently in few aspects so that it can alert the pharmacist of the expiry date of drugs at a given time and handle all departments in the hospital will be an attractive research in future. Apart from this providing best offers and prices of the medicines or tests, providing health tips and video conferencing with the concerned doctor can be further enhancement in the proposed developed system.

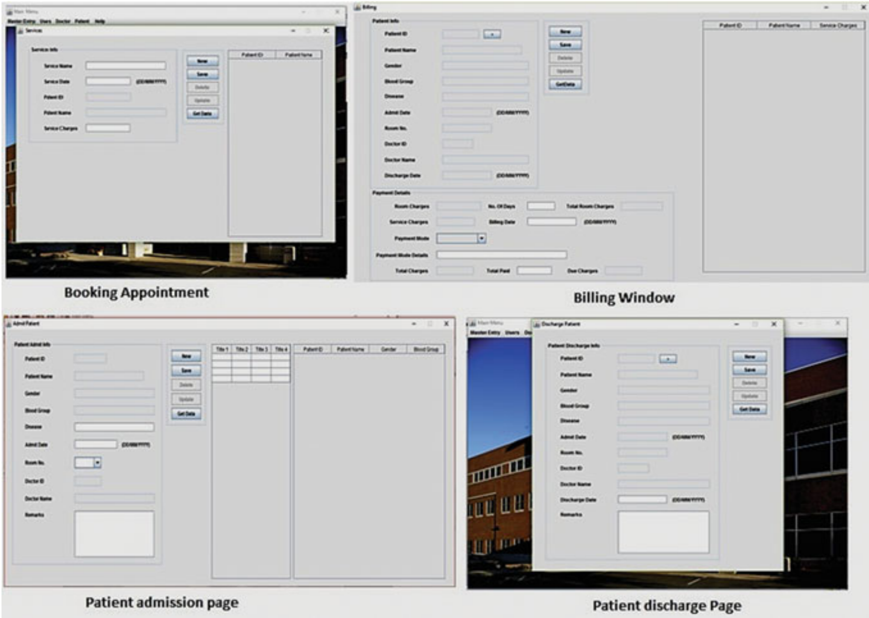


Fig. 3 Appointment booking, billing, patient admission and patient discharge windows

References

1. Ndira, S.P., Rosenberger, K.D., Wetter, T.: Assessment of data quality of and staff satisfaction with an electronic health record system in a developing country (Uganda): a qualitative and quantitative comparative study. Methods Inf. Med. **47**(6), 489–498 (2008). https://doi.org/10.3414/ME0511
2. Ariyaratne, M.H.B.: A private hospital management system. Sri Lanka J. Bio-Med. Inform. (2011). http://dx.doi.org/10.4038/sljbmi.v1i0.3551
3. Buchbinder, S., Thompson, J.: An overview of healthcare management (2010)
4. Praveen, K.A., Gomes, L.A.: A study of the hospital information system (HIS) in the medical records department of a tertiary teaching hospital. J. Acad. Hosp. Adm. (2006)
5. Al-Taee, M.A., Al-Nuaimy, W., Muhsin, Z.J., Al-Ataby, A.: Robot assistant in management of diabetes in children based on the Internet of things. IEEE Internet of Things J. (2016)
6. Wooton, R., Criag, J., Patteson, V. (eds.): Introduction to Telemedicine. The Royal Society and Medicine Press, London (2006)
7. Smith, M., van der Pijl, G.: Developments in Hospital Management and Information Systems. Tilburg University School of Economics, Netherlands (1999)
8. Garrido, T., Raymond, B., Jamieson, L., Liang, L., Wiesenthal, A.: Making the business case for hospital information systems. J. Healthc. Financ. **31**(2), 21–22 (2004)
9. Archer, N., Fevrier-Thomas, U., Lokke, R.C., Mckibbon, K.A., Straus, S.E.: Personal health records. J. Am. Med. Inform. Assoc. **18**(4), 515–522 (2011)
10. OECD report.: Competition in Hospital Services. Directorate for Financial and Enterprise Affairs Competition Committee. Competition Policy Roundtables, Unclassified document-DAF/COMP (2012)

11. Li, Z., Yao, W., Chu, C.-H.: The Use of RFID in healthcare: benefits and barrier. In: 2010 IEEE International Conference on RFID-Technology and Applications (RFID-TA), China (2010). https://doi.org/10.1109/rfid-ta.2010.5529874

Study of Classification Techniques on Medical Datasets

Girish Kumar Singh, Rahul K. Jain and Prabhati Dubey

Abstract Medical science is using digital equipment and generates and gathers large volume of data. These medical datasets are analyzed to get useful information which helps in making decision about diagnosis and treatment. Data mining techniques solve the problem of knowledge extraction from databases from different sources. Several data mining methodologies like Classification, Clustering are used to analyze the data. Classification is a technique used in prediction and to classify the unknown data to a class. This paper presents a study of application of classification algorithms on different kinds of medical datasets.

Keywords Classification · Medical dataset · *k*-neighbor · Neural network SVM

1 Introduction

Techniques for data storing have been changed from primitive file system to relational database system in past decades. Storing data from single data files to different types of databases and data repositories discloses a serious issue which is retrieving of knowledge and information. Data mining techniques solve the problem of knowledge extraction from databases from different sources. Several data mining methodologies like Classification, Clustering are used to analyze the data. These methods provide a relation among data and help us to predict result after analysis. Data mining is a very useful tool in medical field like Disease Prediction, Drug Suggestion, and Treatment

G. K. Singh (✉) · P. Dubey
Department of Computer Science and Applications, Dr. Harisingh Gour University, Sagar, India
e-mail: gkrsingh@gmail.com

P. Dubey
e-mail: prabhatidubey15@gmail.com

R. K. Jain
B.T.I.R.T. College, RGPV University, Bhopal, India
e-mail: rahulkumarjain16@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_57

Effectiveness, etc. Classification techniques are used to classify diseases based on patient's medical history data. This field is required to early detection of dangerous diseases like cancer to provide better treatment and minimize suffer of patient from these diseases. Use of such data mining techniques can save human lives and enhance cure rate of diseases.

1.1 Data Mining

Data mining is a combination of several processes and sub processes. These processes are performed to get meaningful information and knowledge form available data. So data mining is actually information mining or knowledge extraction from large datasets. This information may be helpful to make predictions and decisions. Information extraction techniques of data mining are applied in various fields like Engineering, Science, Medicine, Business, Education, E-commerce, etc. Pattern evaluating techniques are used in most of these fields. Data analysis tasks such as Decision Making, Pattern Evolution are major problems which can be solved by Data mining techniques. In Data mining, different concepts and methods have been used from other fields such as Machine learning, Artificial Intelligence, Statistics, etc.

1.2 Medical Dataset

In the present era of medical science, digital equipments are frequently used. Use of digital equipments generates and gathers large volume of data. These datasets are called medical datasets in which information about patients and measurements of diseases are stored in an organized manner with integrity. These data repositories store data in various attributes values according to pattern of data. Medical datasets can be defined as composition of prepared data about diseases and their effecting conditions.

To make a correct decision for diagnosis and treatment of disease, it is necessary to analyze medical datasets to get useful information which helps in making decision about diagnosis and treatment. Collecting such medical data is not only beneficial to the individual patient but also whole medical society to draw some pattern and diagnosis about disease. For medical field it is essential to understand disease profiles and its impacting pattern to categorized data. Describing what has happened and predicting what will happen is the most useful information that could get from medical datasets. Determining what to do about it is become easy using uncover patterns from vast medical data sources.

1.3 Classification

Classification is a technique used in prediction to classify the unknown data to a class. In classification, data records are classified to reveal a decision, based on its learning models. A model is trained from training datasets to predict the class of unseen data. These training data tuples contain a dependent attribute and a set of predictor attributes. Predictor attributes have some information about data and the value of dependent attribute also called class label is predicted by the learned model. So a categorical label is provided to data tuples by learned model. Better error free efficient prediction without ambiguity is the main objective of any classification technique.

Classification can perform using Supervised or Unsupervised learning methods. Supervised learning trained machine based on derived models which are derived by given training set of tuples and their labels. After successful classification of training tuples, trained machine can be able to classify other unseen tuples. The learning process is unsupervised when the input examples are not class labeled. In which the class label of each training tuple is not known and the number or set of classes to be learned may not be known in advance.

The goodness of classification techniques depends on the accuracy of prediction of unseen data. To analysis the accuracy of the classification technique is accomplished by a part of available data, taken as test data. Accuracy of a classification algorithm can be defined as percentage of correct prediction over total unseen data.

2 Classification Techniques and Its Applications in Medical Science

Today many people are affected by diseases like cancer, liver, heart, kidney, diabetes, dermatologic disease. These diseases are enough complicated to understand and recognized the cause of diseases. Many problems are faced by today's Medical practitioner, Doctor and Drug Researcher to understand the type of disease and which drug should be used and in which amount. It is very necessary to correctly identify diseases to provide treatment and to reduce disease risk. Recognition of actual diseases is very necessary to indentify and treatment of a diseases. Classification is very helpful in indentify a disease. It also helps us to early prediction of diseases and treatment of patient. Among these, classification plays a pivot role to successful implementation of advanced technology for correctly recognition of disease as well as cure.

Advance medical techniques for treatment of these diseases are improved with growing technology. Data mining is emerging field in health care because data mining techniques like classifications, clustering can play an important role to detect and treatment various kinds of diseases.

Biological researchers also use data mining techniques to find new kind of disease and also for a successful cure for new diseases. New diagnoses of diseases in the last century are also relying on advancement of technology. In disease precaution cloud computing is emerging advanced technology in medical field. Cloud computing wearable devices like smart watches are implementing to keep observation on patient health. These devices are useful in analysis and observation of many diseases like heart, lung and diabetic status of patient. They compare risk level of a person with stored data and classify disease and its risk level at any time [1].

Classification is useful in efficiently management of hospital resource. For patient, it is helpful to get knowledge about customer response of a hospital and doctor to find advanced and correct technical treatment of disease in a hospital. Doctors should use various data mining approaches to recognized diseases, patient risks and knowledge about diseases which are difficult to perceive through normal tests. From a doctor's point of view, classification provides a broad area to apply it in medical fields. It is applicable to determination of diseases, selection of drugs, finding an appropriate treatment, Drug risks and Drug quantity for a patient to diagnose effectively. Image pattern recognition is a emerging filed where various CT scan, MRI images are analyzed using clustering and classification.

It is useful if appropriate technique of classification is use to recognize and treatments of a particular type of diseases. To correctly classify a disease many algorithms are available based on some attributes features of disease. These attributes are contained information about patient body organs to diagnosis to disease. To classify a disease, selection of a good technique is very necessary. Selection of techniques depends on its performance and success rate.

Many algorithms have been founded for achieving classification and able to build models. There are many different techniques like Decision Tree Induction [2], Rule Based Classification [3], Support Vector Machine (SVM) [4], Genetic Algorithms [5], Neural Network [6], Rough Set Theory [7], Fuzzy Set Approach [8], k-Nearest Neighbor Classifier [9].

In this study classification techniques such as Decision Tree Induction, Neural Network, SVM and kNN methods have been studied. This study focuses on applicability of these techniques in medical datasets for the classification of disease. In next subsection these classification methods are described.

2.1 Decision Tree Induction

The method of decision tree classification was given by J. Ross Quinlan in 1970–80 [2]. Decision Tree algorithm based on a tree-like structure where tests are performed on each node of tree. In decision tree, each internal node represents a test on an attribute and each branch of the node depicts the outcome of the test. Every terminal node shows a class label for testing tuple. The output is defined by terminal (leaf) node. Decision tree is constructed in top-down recursive manner. ID3 (Iterative Decotomiser), C4.5 and CART (Classification and Regression Tree) are three algorithms mostly used to construct a Decision Tree. Decision Tree algorithm develops a learning model to determine the value on each node to perform testing. In testing, value of given tuple is compared with the value of node obtained by some methods, according the result of test a class label is assigned for tuple.

It is necessary to select some appropriate attributes for testing to find a correct splitting criterion. There are also some methods to select appropriate attribute from total number of available attributes. To perform this task, three popular methods are Information Gain [10], Gain ratio [11] and Gini Index [12] used are used. Pre-pruning and Post-pruning are two methods to prone a decision tree. The pruning method is used to eliminate anomaly it occurred in tree. The time complexity of Decision Tree Induction Method is $\mathbf{O}(\mathrm{m} \cdot \mathrm{n}^2)$, where m is the size of training data and n is number of attributes [13].

2.2 Support Vector Machine

The concept of support vector machine (SVM) was proposed by Vladimir Vapnik, Bernhard Boser and Isabelle Guyon in 1992 [4]. Statistical learning theory is considered as basis work for SVM. This work was done by Vladimir Vapnik and Alexei Chervonenkis in 1960. SVM can be used for both classification and prediction. The concept behind SVM is a decision boundary that is defined in an imagined surface. A Support Vector Machine method can classify tuples those have two-dimensional or multi-dimensional attribute values. Support vector machine algorithm draws a surface that defines a boundary among many data tuples which represent examples plotted in multi-dimensional space according their attribute (feature) values. The main objective of SVM based classification is to create a decision boundary, called a hyperplane that able to partition of data tuples on either side based on tuples attributes values. It categorized data tuples which related to same class through a hyperplane and separate them in different regions. A hyperplane can be found by using support vectors and margins. Support vectors are essential training tuples and define margins. Hyperplane is defined as minimum distance between support vectors. Any training tuples that fall on hyperplane are called support vector. An outer boundary is made by data tuples on each side. These outer boundaries are referred as Convex Hull. A hyperplane basic equation is

$$W \cdot X + b = 0,$$

where $W = (W_1, W_2, \ldots, W_n)$ is a weight vector, where n is the number of attributes and b is a scalar and is referred as a bias.

2.3 k-Nearest-Neighbor Classifiers

k-nearest-neighbor was described in 1950 [9]. This algorithm classifies training tuples on analogy based learning. It classifies the training tuples on basis of similarities and comparison to a particular type. To classify an unknown tuple, it finds the pattern space that is closest to the unknown tuple. Pattern space is defined by some existing training data tuples. If there is k training tuples, these k training tuples are nearest neighbor for a given unseen tuple. This algorithm finds a nearest neighbor for an unseen tuple among giving these k-neighbors. Then the unknown tuple is classify according a nearest neighbor tuple value.

To determine the closeness of unseen tuple with respect to giving k neighbor set of tuples distance matrix is used like Euclidean Distance Matrix. The Euclidean Distance can define as distance between two points or tuples.

$$\mathrm{dist}(X_1, X_2) = (\sum_{i=1}^{n}(x_{1i} - x_{2i})^2)^{1/2},$$

where $X_1 = (x_{11}, x_{12}, \ldots, x_{1n})$ and $X_2 = (x_{21}, x_{22}, \ldots, x_{2n})$ represent tuples and their numeric attributes. If there are n numbers of samples then the time complexity of KNN is **O**(n) [14].

2.4 Neural Network

Neural network is a classification and prediction technique based on concept of biological system represents a brain image or symbol. It has biological system structure like brains neurons structure. Neural network has the great ability to prediction, forecasting and classification [6].

Neural Network is a multilayer perception technique that trains model by using Backpropagation method. A neural network is a set of connected input/output units in which weight is associated with each connection. These weights are defined according input values. Inputs are fed simultaneously into the input layer of network. Input units value correspond to attributes values of training tuples. This input is transferred to second layer in network called hidden layer after adjusting some weight upon inputs. There may be a number of hidden layers in network. The output of one hidden layer can be input to a next hidden layer. Finally an output layer represents the class labels of tuples. The associated weights upon hidden layer decide the output for a tuple (its class label).

Neural network algorithms are successfully useful on a wide verity of real word data including hand writer charterer recognition, pathology and laboratory medium and training a computer to pronouns English. Time complexity of The time complexity of ANN is **O**(nlogn) [15].

Table 1 Accuracy prediction of diseases in percentage

Accuracy of techniques				
Datasets	SVM (%)	Neural network (%)	Decision tree (%)	k-nearest neighbor (%)
Blood transfusion	72	72.66	72.66	70.66
Breast cancer	91.43	91.42	92.86	84.29
Diabetic retinopathy debrecen	74.35	70.43	60.87	66.96
Statlog (Heart)	76.79	66.07	64.29	60.71
Dermatology	98.67	82.67	90.67	88

3 Result and Discussion

In this study five datasets available at UCI Database repository [16] namely Blood Transfusion Service Center Data Set, Breast Cancer Wisconsin (Diagnostic) Data Set, Diabetic Retinopathy Debrecen Data Set, Statlog (Heart) Data Set, Dermatology Data Set has been used to compare four well known techniques for classification. Four classification techniques namely SVM, ANN, Decision Tree and KNN are performed on these five disease datasets using MATLAB Software.

Blood Transfusion Service Center dataset is provided by Blood Transfusion Service Center in Hsin-Chu City in Taiwan. This Multivariate Dataset has four attributes for classification and one attribute for the deciding whether donor can donate or not. It has 748 instances.

Breast Cancer Wisconsin (Diagnostic) Dataset is created in University of Wisconsin. Features are determined by a digitized image of a fine needle aspirate (FNA) of a breast mass. The 10 real-valued features are computed for each cell. The number of samples is 569 in this dataset.

Diabetic Retinopathy Disease is a Diabetes complication that affects eyes. Diabetic Retinopathy Dataset is created by Dr. Balint Antal and Dr. Andras Hajdu in university of Debrecen, Hungary. It has 18 attributes and 1151 number of instances.

Statlog (Heart) data set stands for Statistical and Logical Learning Algorithm. This process used to determine presence and absence of heart disease in a person. This database contains 13 attributes. It has a resulting attribute with class label 1 and 2. The number of instances is 270 in this dataset.

Dermatology Data Set is created by Dr. Nilsel Ilter and Dr. H. Altay Guvenir. It has 33 attributes and 366 instances. Prediction of instances will be in 4° of ranges where 1 indicates that the feature is not present, 4 indicate the largest possible amount, 2 and 3 indicate the relative intermediate values.

Result of four classification techniques on these datasets is presented and compared in terms of percentage of accuracy as shown in Table 1.

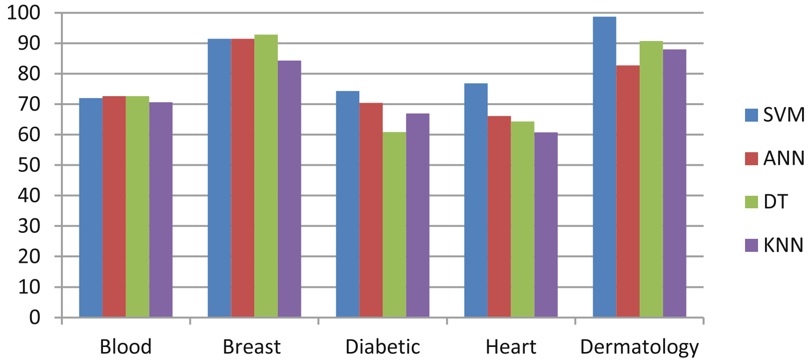


Fig. 1 Percentage classification of different dataset using various algorithms

The given figures are depicting the result of classification of diseases datasets according classification techniques. On the basis of accuracy, an appropriate classification technique can be selected for a particular disease.

Results of the classification techniques on various datasets are compared in Fig. 1.

The accuracy of all the four techniques is almost same for Blood Transfusion dataset with more 70% accuracy. For Breast Cancer dataset SVM, ANN and Decision Tree gives good results with more than 91% accuracy. KNN gives 84.29% accuracy. Decision Tree gives the highest accuracy while SVM and ANN give almost same result. SVM gives 74.35% accuracy which is highest among for Diabetic dataset while ANN shows 70.43% accuracy. For the same dataset Decision Tree and KNN gives accuracy less than 67%. The result of SVM very impressive compared to other techniques for both (Statlog) heart disease dataset and Dermatology dataset.

4 Conclusion

It is very useful for humankind to track disease and provide significant cure for diseases. Use of previous data history of disease is an emerging technology for cure and treatment. It is helpful in early detection and treatment of diseases. In this paper, the use of four algorithms has been demonstrated on medical dataset. After analysis of all disease and algorithms, it has been found that the performance of SVM algorithm better than or equivalent to other techniques. SVM can be used for many datasets on the basis of good performance. Classification technique which has shown the highest accuracy rate over the datasets may be selected the classification technique for medical science.

References

1. Vanaja, S., Rameshkumar, K.: Performance analysis of classification algorithms on medical diagnoses-a survey. J. Comput. Sci. 30–52 (2014)
2. Quinlan, J.R.: Induction of Decision trees, pp. 81–106. Kluwer Academic Publishers (1986)
3. Kotsiantis, S., Kanellopoulos, D.: Association rules mining. A recent overview. Int. Trans. Comput. Sci. Eng. 71–82 (2006)
4. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, New York, USA, pp. 144–152 (1992)
5. Eiben, A.E., et al.: Genetic algorithms with multi-parents recombination. In: The Third Conference on Parallel Problem Solving from Nature, pp. 78–87 (1994)
6. Hopfield, J.J.: Artificial neural networks. IEEE Circuit Device Mag. 3–10 (1988)
7. Pawlak, Z.: Rough set. Int. J. Comput. Inf. Sci. 341–356 (1982)
8. Zadeh, L.A.: Fuzzy Sets, pp. 338–353. Elsevier (1965)
9. Altman, N.M.: An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat. 175–185 (1992)
10. De Mántaras, R.L.: A distance-based attribute selection measure for decision tree induction, pp. 81–92. Kluwer Academic Publishers-Plenum Publishers (1991)
11. Prasad, N.: Gain ratio as attribute selection measure in elegant decision tree to predict precipitation. In: Modelling and Simulation (EUROSIM), pp. 141–150 (2008)
12. Zhang, S., et al.: A strategy for attributes selection in cost-sensitive decision trees induction. In: Computer and Information Technology Workshops, pp. 8–13 (2008)
13. Su, J., Zhang, H.: A fast decision tree learning algorithm. In: Proceedings of AAAI'06 Proceedings of the 21st National Conference on Artificial Intelligence, vol. 1, pp. 500–505 (2006)
14. Ismanto, H., Wardoyo, R.: Analysis of C4.5 and K-nearest neighbor (KNN) method on algorithm of clustering for deciding mainstay area. IOSR J. Comput. Eng. 86–92 (2016)
15. Orponene, P.: Computational complexity of networks: a survey. Nordic J. Comput. 94–110 (1994)
16. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets.html

Feature Ensemble Learning Based on Sparse Autoencoders for Diagnosis of Parkinson's Disease

Vinod J. Kadam and Shivajirao M. Jadhav

Abstract Parkinson's disease detection through proper representation of the vocal and speech datasets remains an important classification problem. For this problem, we proposed a feature ensemble learning method based on sparse autoencoders. The dataset for this purpose was obtained from UCI, an online repository of comprehensive datasets. Some simulations were conducted over the UCI dataset to confirm the effectiveness of the proposed model. In this paper, the outcomes of the experimentation are compared with the outcomes of stacked sparse Autoencoders and softmax classifier based deep neural network and many classification techniques. Our proposed method yields superior results than DNN. With the proposed model, we obtained a true promising accuracy more than 90%. The outcome of the study also proves that the Feature ensemble learning based on sparse autoencoders method is comparable to other methods present in the literature. The experimental results and statistical analyses are pointing out that the proposed classifier is really useful and practical model for Parkinson's disease investigation.

Keywords Parkinson's disease · Stacked sparse autoencoders
Softmax classifier · Ensemble learning

1 Introduction

1.1 Parkinson's Disease

Parkinson's disease is one of the main health-related issues all around the globe. It is the prototypic adult-onset neuro-degenerative disorder initially described by

V. J. Kadam (✉) · S. M. Jadhav
Department of IT, Babasaheb Ambedkar Technological University,
Lonere, Raigad, Maharashtra, India
e-mail: vjkadam@dbatu.ac.in

S. M. Jadhav
e-mail: smjadhav@dbatu.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_58

Doctor James Parkinson as shaking palsy [1]. It is a chronic disorder and condition of patient progressively become worse. In the biological neural network, a chemical called Dopamine is released by neurons and it is required to govern many motor and non-motor natural actions of the human body. In Parkinson's disease, certain groups of cell bodies of neurons are not able to generate Dopamine. As Parkinson's disease advances, the quantity of Dopamine generated in the nervous system reduces, leaving a human being unable to manage movement normally [2, 3]. Parkinson's disease usually begins with a tremor in one hand, foot, or leg. Additional indications are a slow movement, stiffness, reduced body balance, trouble in standing, decreased facial expressions, difficulties with coordination, trouble in thinking, difficulty in understanding, difficulty to write, distorted sense of smell, dribbling of urine, impaired voice, soft speech, and voice box spasms etc. [2–6] 90% of these patients have vocal impairment and troubles in talking or speaking. Schley et al. [7] Hence, careful examination of the sound/voice of the people with Parkinson's disease by superior signal processing procedures helps in the diagnosis and supports in the tracking of the progress of this disease.

1.2 Related Work

Due to advancement in the technology, new tools for various diseases identification and detection become possible. Many machine learning, expert systems and soft computing methods are proposed by various scholars in the almost every sphere of the medical system. Various methods used in literature for diagnosis of Parkinson's disease through proper representation of the vocal and speech datasets. Important steps of these computerized classification systems are prepossessing, feature extraction and classification. Little et al. [8] applied kernel support vector machine (SVM) with feature selection method to the diagnosis of Parkinson's disease, they obtained promising true accuracy of 91.4%. Little et al. [8], Shahbaba and Neal [9] proposed Nonlinear Models using Dirichlet Process Mixtures for Parkinson's disease diagnosis. Shahbaba and Neal [9], Psorakis et al. [10] used Improved multi-class multi-kernel Relevance Vector Machines (mRVMs) and with ten-fold cross-validation method, they obtained 89.47% accuracy. Psorakis et al. [10], Guo et al. [11] proposed a model with genetic programming and expectation-maximization algorithm. Guo et al. Psorakis et al. [11], Sakar and Olcay [12] used the mutual information measure and Support Vector Machines for building a suitable model. Sakar and Kursun [12], Das [13] carried out a comparative study of four independent classification models and obtained the highest accuracy 92.9% from the Artifical Neural Network-based model. Das [13], Luukka [14] applied fuzzy entropy measures and similarity classifier. With 50% training and 50% testing set, this study obtained average true accuracy 85.03%. Luukka [14], Ozcift and Gulten [15] proposed a correlation based feature selection (CFS) approach with rotation forest ensemble classifiers for diagnosis of Parkinson's disease. Ozcift and Gulten [15], Li et al. [16] proposed fuzzy-based non-linear transformation techniques with PCA and SVM to increase the performance of classification on relatively small data. They applied this method to six medical datasets including Parkinson's disease dataset to illustrate the performance. Li et al. [16], Åström and Koker [17] used a Parallel Artificial Neural Network model and obtained 91.20% accuracy. Åström and Koker [17], Spadoto et al. [18] applied evolutionary based feature selection techniques in order to improve the accuracy of optimum path forest classifier for detection of Parkinson's disease. Spadoto et al. [18], Polat [19] applied Fuzzy c-means(FCM) clustering-based feature weighting method and adopted kNN classifier and obtained 97.93% encouraging classification accuracy. Polat [19], Daliri [20] applied Support Vector Machine with chi-square distance kernel. The accuracy obtained was 91.20%. Daliri [20], Chen et al. [21] used PCA and fuzzy kNN system. They achieved the promising accuracy of 96.07% with 10 fold cross-validation method. Chen et al. [21], Zuo et al. [22] applied an evolutionary technique (Particle Swarm Optimization) to enhance the performance of fuzzy k-nearest neighbor classifier and obtained the best accuracy of 97.47%. Zuo et al. [22], Zhang [23] described diagnosis of Parkinson's disease using time-frequency features, stacked autoencoders (SAE), and kNN classifier Zhang [23].

1.3 Motivation and This Work

Recently, deep neural networks have verified to be the superior method not only for various computer vision but also for other classification applications. In conventional shallow classifiers, feature engineering is important to improve performance, but the Deep Neural Networks automatically learn hierarchies of relevant features directly from the training data. Therefore, Deep Neural Networks (DNNs) may extend a potentially superior classifier for the speech of the people with Parkinsons disease over the conventional methods. Deep Neural networks are well known for their strong function approximations. Unlike common classification techniques, DNNs performs both dimension reduction and classification. In DNNs, Stacked sparse autoencoders decrease the dimension of the features and Softmax layer classify the representations. The Training of a DNN requires greedy layer-wise pre-training phase and fine-tuning phase. Pre-training phase is intended to initialize a Deep Neural Network. It is performed through unsupervised feature learning, with many feature representations produced. In DNNs, The outermost layer representation gives more abstraction and compactness, therefore, this representation is used to train one Softmax classifier. Proposed a feature ensemble learning model make use of all layers representations for Parkinsons disease classification. In this approach, three Softmax classifiers are trained by using different representations. Y. Lu, L. Zhang, B. Wang and J. Yang proposed feature ensemble learning based on Sparse Autoencoders for image classification to improve performance [24]. Motivated by these developments, we propose this feature ensemble learning based on sparse autoencoders for the automated diagnosis of the Parkinsons disease. The effectiveness of the proposed model to classify the speech impairments is evaluated on one real Parkinson's disease dataset available on UCI machine learning repository [25]. Experiment outcomes prove that applying the proposed model for classification of that dataset gives the better result. The experimental results also intimate that the proposed model is an efficient alternative for the diagnosis and classification of the Parkinsons disease.

The remaining part of the paper is arranged as follows. Section 2 reviews Stacked Autoencoders and Softmax layer based Deep Neural Network Classifier. In Sect. 3, a brief description of the proposed Feature ensemble learning based on sparse autoencoders for Parkinson's disease diagnosis is given. Section 4 demonstrates the experiments and results. Section 4 also includes the dataset, performance indices and various comparisons.

2 Deep Neural Network

Deep learning methods arise as an extremely powerful scheme. The main advantage is that they have an arrangement that enables extracting features from given dataset without any extra prepossessing. The DNN consists of many simple structures that are arranged to develop a stack. All of these simple structures do non-linear operations, adjusting the data size to reproduce the data in a distinct space, helping to show hidden features in the given dataset.

2.1 Sparse Autoencoders

The Autoencoder is a three-layered feedforward artificial neural network. It tries to produce its input as the output of the network that may generate another representation of inputs. The main objective of Autoencoder learning is to minimize an average distortion between input vectors and output vectors. It is an artificial neural network for coding based on a two principal 1. A common method of training called backward propagation of errors 2. Target output is equal to input i.e. sizes of the input layer and the output layer are same. It relies on the concept of approximation to the identity function. It comprises two parts: the encoder and decoder. It is a symmetrical neural network. Encoder? part maps high-dimensional input data to its low dimensional encoded representation and the decoder? part takes this low dimensional encoded representation and generates original high-dimensional data. An approximation to the identity function sounds trivial in learning problem but the use of fewer number neurons in hidden layer than that of input layer makes it more interesting. This makes it under complete Autoencoder where interest is not in the output of the network. In this way, network learns compressed data instead of identity function. Commonly used activation functions for nonlinear mapping in Autoencoder are sigmoid activation function and rectified linear unit. If we use linear activation function, Autoencoder will be similar to principal component analysis. The cost function for training a Sparse Autoencoder is given by Eq. 1.This cost function includes 3 terms. The first term is mean square error which gives the discrepancy between input x and reconstructed $\hat{x}$ over the whole training dataset [26].

$$\begin{aligned} E &= MeanSquaredError \\ &+\lambda \times L2RegularizationTerm \\ &+\beta \times SparsityRegularizationTerm \ , \end{aligned} \tag{1}$$

where λ is the coefficient for the L2 regularization term and β is the coefficient for the sparsity regularization term. L2 regularization term is given by Eq. 2.

$$L2RegularizationTerm = \frac{1}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}\left(w_{ji}^{(l)}\right)^2 \ , \tag{2}$$

where n_l = Number of layers, l = Layer l, s_l = Number of units in l layer, $w_{ji}^{(l)}$ = the weight value between node i in the layer l and node j in layer $l+1$ Sparsity Regularization Term is defined as

$$SparsityRegularizationTerm = \sum_{j=1}^{s_2} KL(\rho \| \hat{\rho}_j) \ , \tag{3}$$

where $\hat{\rho}_j$ =average activation of hidden node j, ρ is a sparsity parameter.

$KL(\rho \parallel \hat{\rho}_j)$ is Kullback-Leibler divergence is defined as

$$KL(\rho \| \hat{\rho}_j) = (\rho)log\left(\frac{\rho}{\hat{\rho}_j}\right) + (1-\rho)log\left(\frac{1-\rho}{1-\hat{\rho}_j}\right) \ , \tag{4}$$

An autoencoder can also be fine-tuned to make the learned representation more suitable for classification using labeled data after the unsupervised stage is finished by replacing its decoder layer with an output layer for label prediction.

2.2 Softmax Regression

Logistic regression is a binary linear classifier and another log-linear model called Softmax regression is multinomial logistic regression. The softmax is the superior variant of the logistic regression. It is a generalization of logistic regression for the multi-category problem. A softmax classifier which is trained on given training dataset gives output a separate probability for each of the category, and the probabilities all add up to 1. The softmax is attached to the stacked Autoencoders as the last supervised learning layer which is used to classify the features given by stacked Autoencoders.

2.3 Stacked Sparse Autoencoders and Softmax Regression as Classifier

A special type of Deep learning network called Stacked sparse Autoencoder is a hierarchical feature learning approach. In this DNN, there are multiple layers of Autoencoders and output of each layer is given as input to successive layer. The unsupervised training process, also called pre-training? benefits from the immediate availability of a reconstruction error which is a distance between the output and the input. Once a layer correctly reconstructs its input, another layer can be added or stacked on top of it and pre-trained to do the same with the hidden layer or the reconstructed output of the previous layer as input. The final Softmax layer representing a log-linear classifier makes use of the most abstract feature vectors represented in the topmost hidden layer. Once the overall architecture is trained in this way, it can be fine-tuned for achieving a more specific task.

3 Proposed Feature Ensemble Learning Based on Sparse Autoencoders for Parkinson's Disease Diagnosis

In this approach, we trained two sparse Autoencoders SAE1 and SAE2 so as two get two separate representations of given data. The output of hidden layer of SAE1 is given as input to SAE2 to get the second representation of data. The first and second representations are then concatenated to get the third representation of given data. The three Softmax classifiers (SM1, SM2, and SM3) are employed for classifications using above mentioned three representations. For first two classifiers, we need to perform fine tuning including the input layer, feature representation layer, and Softmax layer to improve the performance of the Softmax layers SM1 and SM2. Figure 1 shows the proposed architecture. Training steps of the whole network are given below (see Fig. 1).

- Step 1
  - Train SAE1 with HS1 hidden units on training dataset x. Hidden layer of SAE1 transforms x into features set FS1.
  - Train SAE2 with HS2 hidden units on training feature set FS1. Hidden layer of SAE2 transforms FS1 into features set FS2.
  - Train SM2 on the features set FS2.
  - Fine tune whole network SAE1, SAE2, and SM2.
- Step 2
  - Apply input data x to ASE1 to get the feature set FS1.
  - Train SM1 on the features set FS1.
  - Fine tune whole shallow network SAE1 and SM1.

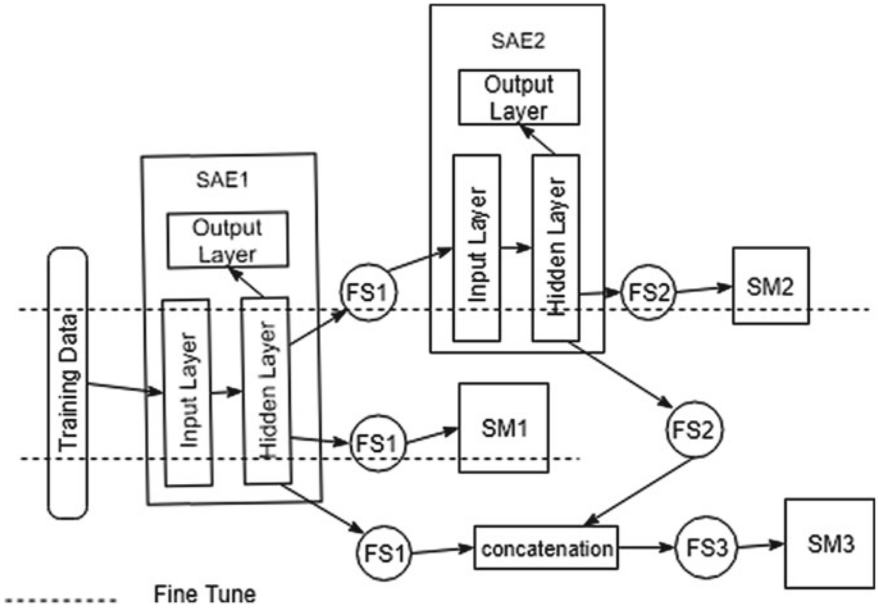


Fig. 1 Overview of training phase of proposed (Feature Ensemble Learning Based on Sparse Autoencoders) method. (SAE-Sparse Autoencoder, FS-Feature Space, SM-Softmax classifier)

- Step 3
  - Apply input data x to ASE1 to get feature set FS1
  - Apply feature set FS1 to ASE2 to get the feature set FS2.
  - concatenate feature set FS1 and feature set FS2 to get the feature set FS3.
  - Train SM3 on the features set FS3.

For testing phase, to combining the votes of three classifiers, we considered Naive Bayes combination method. Naive Bayes combination method assumes that the classifiers are mutually independent. Naive Bayes combination methods are MAX rule, MIN rule and AVG rule etc. Details are given by [27, 28]. In our Experimentation, we considered AVG rule Naive Bayes combination method. Suppose x is new given sample for testing, its label y takes on different possible values $j = 1, 2, \ldots k$, we can get corresponding prediction probabilities for the softmax classifier SM_n ($n = 1, 2, 3$ here), denoted as $P_{nj}(x)$. The final value of label y is determined by the following rule

$$y = arg \max_{j \in \{1,2,\ldots k\}} \sum_{n=1}^{N} P_{nj}(x)/N \ , \tag{5}$$

where N = 3 and k = 2 here. Figure 2 shows the testing phase of proposed method.

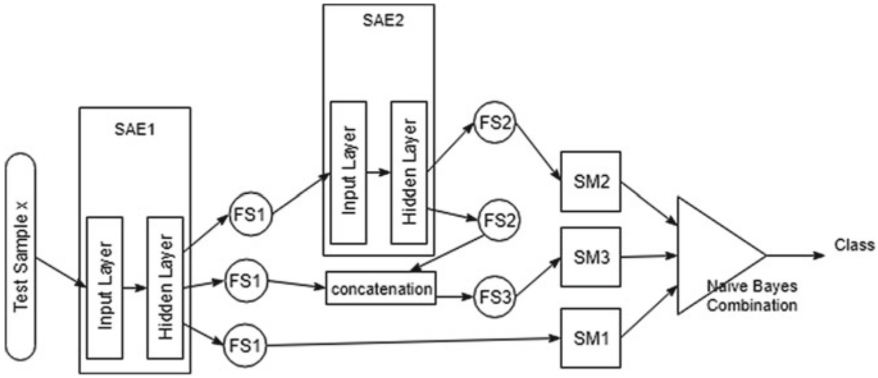


Fig. 2 Overview of testing phase of proposed (Feature Ensemble Learning Based on Sparse Autoencoders) method. (SAE-Sparse Autoencoder, FS-Feature Space, SM-Softmax classifier)

4 Experiments and Results

4.1 Dataset

The main aim of the experimentation is to classify healthy people and people with Parkinson's disease (PWD). The UCI dataset named "Oxford Parkinson's Disease Detection (OPD)" of Biomedical voice measurements is used for this purpose [8, 29]. Details about the OPD dataset are given in the Table 1

Table 1 UCI dataset: Oxford Parkinson's disease detection [25]

Data set characteristics:	Multivariate
Attribute characteristics:	Real
Number of attributes:	22
Number of instances:	195
subjects involved	31
Healthy control subjects involved	8 (5 females and 3 females) averaged age of 60.2 year
PWD[a] involved:	23 (7 females and 16 males) averaged age of 67.8 year
Classification type:	Binary (healthy (0) and PWD[a] (1))
Voice recordings per person:	about 6
Missing value:	No
Generated by:	Max A. Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado

[a]people with Parkinson's disease

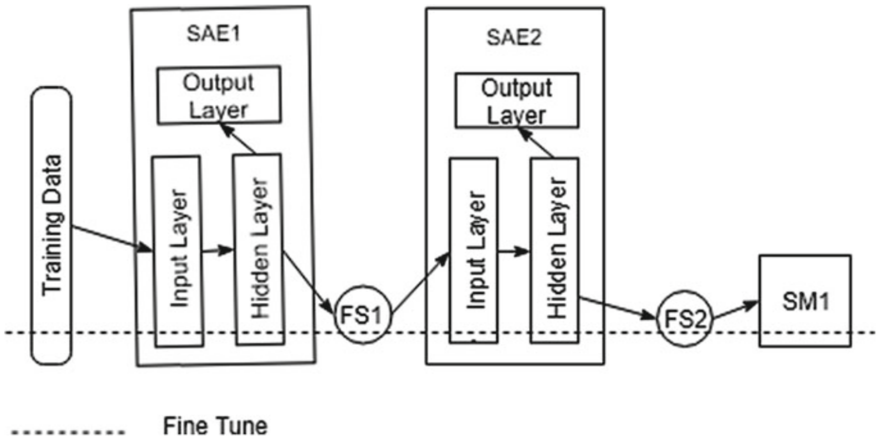


Fig. 3 Proposed Deep Neural Network model (SAE-Sparse Autoencoder, FS-Feature Space, SM-Softmax classifier)

4.2 Experimentation

To compare the performance of proposed Feature ensemble learning based on Sparse Autoencoders model (henceforth referred to as FESA-DNN) over OPD dataset, we also proposed a Stacked Sparse Autoencoders based Deep Neural Network model (henceforth referred to as DNN) as described in Sect. 2. The training and testing procedures of DNN are also given in Sect. 2. Both FESA-DNN and DNN contain two sparse Autoencoders (SAE1 and SAE2). The proposed architecture of DNN (Stacked Sparse Autoencoders with last Softmax layer) is shown in the Fig. 3.

Regulation Parameters For both DNN and FESA-DNN, the activation function was set to sigmoidal. To find the optimal value of parameters λ, β and ρ, we used 50% training data-50% testing data method. Hidden layer size of SAE1 and SAE2 was set as 4 each for both DNN and FESA-DNN. Scaled conjugate gradient with 600 iterations was used for training of all Autoencoders and all softmax classifiers and fine tuning in both of the models. Difference values were assigned to regulation parameters λ, β and ρ to study their effect on accuracy. Coefficient for the L2 regularization term (λ) in the range of 0.003 to 0.09, Sparsity Proportion (ρ) in the range of 0.05–1 and Coefficient for the Sparsity regularization term (β) in the range of 1 to 10 were considered for experimentation. Each individual combination of regulation parameters on DNN and FESA-DNN was tested 10 times with different 50% training data-50% testing data and averaged accuracy is considered for comparison. In 97% cases, FESA-DNN gave higher accuracy than DNN. We obtained top two performances of FESA-DNN at L1:($\lambda = 0.03$, $\rho = 0.5$, $\beta = 8$) and L2:($\lambda = 0.07$, $\rho = 0.05$, $\beta = 2$). We obtained top two performances of DNN at L3:($\lambda = 0.009$, $\rho = 0.2$, $\beta = 3$) and L4:($\lambda = 0.01$, $\rho = 0.8$, $\beta = 9$). Rest of the experiment was conducted with these four parameter settings.

Experiment For experimentation purpose, we considered 8–6, 6–8, 8–8, 8–10, 10–8 hidden layers neurons in SAE1 and SAE2. Scaled conjugate gradient with 400 iterations was used for training of all Autoencoders and all softmax classifiers and fine tuning in both of the models. Parameter settings L1, L2, L3, and L4 were considered. For each individual combination of hidden layer neurons of SAE1-SAE2 and parameter settings, we conducted 10 runs of 10 fold cross-validation method. The averaged result of ten runs of 10 fold cross-validations on each combination were used for comparison. Performance indices used for comparison are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\%; , \tag{6}$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\%; , \tag{7}$$

$$Specificity = \frac{TN}{FP + TN} \times 100\%; , \tag{8}$$

where TP = true positive classifications, FP = false positive classifications , TN = true negative classifications, and FN = false negative classifications and these values are defined in Table 2.

Comparison between FESA-DNN and DNN is given in Table 3 and Fig. 4. Table 4 shows comparison between proposed models and other methods present in the literature.

Discussion In this study, we proposed a Feature ensemble learning method based on Sparse Autoencoders to classify healthy people and people with Parkinson's disease (PWD) through proper representation of the vocal and speech datasets. The UCI dataset Oxford Parkinson's Disease Detection (OPD)? of biomedical voice measurements was used in this study. Ten runs of tenfold cross-validation method was adopted to estimate the accuracy. The averaged results of ten runs of 10 fold cross-validations were used for comparison. Results show that the highest accuracy is achieved by proposed Feature ensemble learning based on Sparse Autoencoders method at setting L1 and hidden layer size 8 and 10 of Sparse Autoencoder 1 and Sparse Autoencoder 2 respectively. The highest accuracy 92.19% is achieved by Stacked Autoencoder

Table 2 Confusion matrix of classification

	Prediction as PWD	Prediction as healthy subject
Actual PWD	TP	FN
Actual healthy subject	FP	TN

Table 3 Comparison between FESA-DNN and DNN (Avg. of ten runs of 10 fold-CV)

Hidden layer size of SAE1 & SAE2	Parameter setting	Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
6–8	L1	FESA-DNN DNN	**92.48** 91.63	95.14 93.19	84.15 **86.75**
6–8	L2	FESA-DNN DNN	91.54 90.61	94.53 92.00	82.45 86.35
6–8	L3	FESA-DNN DNN	90.85 89.95	**95.31** 93.31	77.00 79.45
6–8	L4	FESA-DNN DNN	91.15 90.01	95.30 92.78	78.55 81.65
8–6	L1	FESA-DNN DNN	**92.32** 91.51	94.89 92.38	84.50 **88.80**
8–6	L2	FESA-DNN DNN	91.69 91.23	94.07 92.99	84.35 85.90
8–6	L3	FESA-DNN DNN	90.99 88.50	94.93 92.21	78.00 77.25
8–6	L4	FESA-DNN DNN	91.15 89.32	**95.31** 92.86	78.10 78.35
8–8	L1	FESA-DNN DNN	**93.33** 90.33	94.61 93.28	**89.50** 82.00
8–8	L2	FESA-DNN DNN	91.26 91.16	93.78 92.50	83.55 87.10
8–8	L3	FESA-DNN DNN	90.71 86.77	**95.07** 91.67	77.50 71.75
8–8	L4	FESA-DNN DNN	92.07 88.62	95.02 92.20	83.05 77.05
8–10	L1	FESA-DNN DNN	**93.84** 91.73	95.23 92.47	**90.00** **90.00**
8–10	L2	FESA-DNN DNN	92.01 91.23	**95.49** 92.59	84.45 87.20
8–10	L3	FESA-DNN DNN	90.02 86.60	94.40 93.28	76.50 66.45
8–10	L4	FESA-DNN DNN	91.49 88.20	95.03 93.49	80.85 72.35
10–8	L1	FESA-DNN DNN	**93.80** 91.28	**97.28** 93.28	83.50 85.50
10–8	L2	FESA-DNN DNN	92.33 92.19	94.13 93.59	86.70 **87.75**
10–8	L3	FESA-DNN DNN	89.51 85.29	93.60 91.87	77.10 65.25
10–8	L4	FESA-DNN DNN	90.86 86.85	94.92 92.80	78.60 68.30

Table 4 Comparison with other methods

Study	Method	Test method	Accuracy (%)
Max Little et al. [8]	Pre-selection filter + exhaustive search + support vector machine	Bootstrap with 50 replicates	91.40
Shahbaba and Neal [9]	Dirichlet process mixtures	Fivefold CV	87.7
Psorakis et al. [10]	Improved multi-class multi-kernel relevance vector machines (mRVMs)	10 fold CV	89.47
Guo et al. [11]	Genetic programming and the expectation maximization algorithm	10 fold CV	93.10
Sakar and Kursun [12]	Mutual information based feature selection + support vector machines	Bootstrap with 50 replicates	92.75
Das [13]	ANN decision tree	Hold-out Hold-out	92.90 84.30
Ozcift and Gulten [15]	Correlation based feature selection+ rotation forest (RF) ensemble classifiers	10 fold CV	87.1
Luukka [14]	Fuzzy entropy measures with similarity classifier	Hold-out	85.03
Li et al. [16]	Fuzzy-based nonlinear transformation + support vector machines	Hold-out	93.47
Åström, Koker [17]	Parallel neural network	Hold-out	91.20
Spadoto et al. [18]	Particle swarm optimization + optimum path forest classifier Harmony memory + Optimum path forest classifier Gravitational search algorithm + optimum path forest classifier	Hold-out Hold-out Hold-out	73.50 84.01 84.01
Polat [19]	Fuzzy c-means (FCM) clustering-based feature weighting (FCMFW) +k -nearest neighbour classifier	Hold-out	97.93
Daliri [20]	Support vector machines with chi-square distance kernel	Hold-out	91.20
Chen et al. [21]	PCA-fuzzy k-nearest neighbor	10 fold CV	96.07
Zuo et al. [22]	Particle swarm optimization enhanced fuzzy k-nearest neighbor	10 fold CV	97.47
Zhang [23]	SAE and KNN classifier	Hold-out	94–98
This study	Stacked sparse autoencoders and softmax based DNN	10 fold CV*	92.19
This study	Feature ensemble learning based on sparse autoencoders (FESA-DNN)	10 fold CV[a]	93.84

[a] Average of ten runs of 10 fold-CV

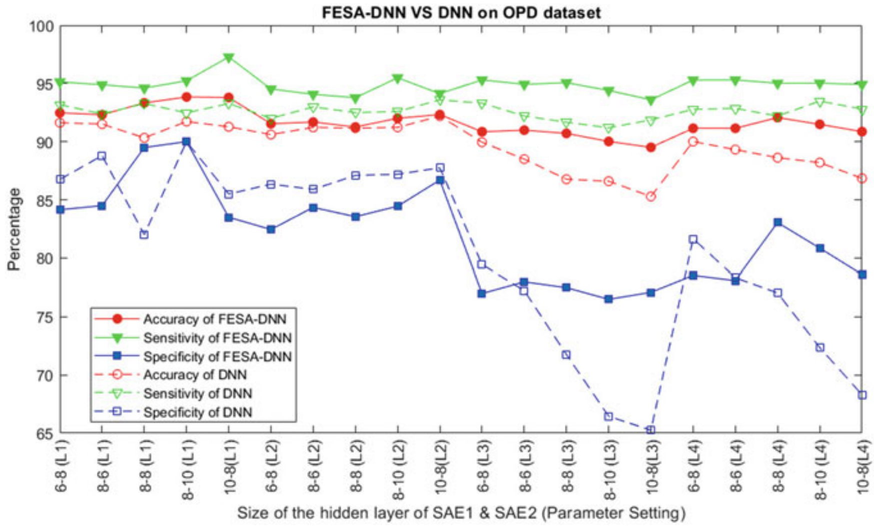


Fig. 4 FESA-DNN versus DNN on OPD dataset (Average of ten runs of 10 fold-CV)

and Softmax based DNN model at parameter setting L2 and hidden layer size 10 and 8 of Sparse Autoencoder 1 and Sparse Autoencoder 2 respectively. Highest Sensitivity and Specificity achieved by Feature ensemble learning based on Sparse Autoencoders method are 97.28% and 90% respectively. Highest Sensitivity and Specificity achieved by DNN method are 93.59% and 90% respectively. Our proposed method yields superior results than DNN. Comparison between FESA-DNN and DNN with other techniques available in the literature is also provided in this paper. FESA-DNN produced the better result than many methods. It produced the comparable and satisfactory result using proper tuning parameter settings.

5 Conclusion

With speedy developments taking place in the domain of biomedical, computer-based automated decision support systems perform a more significant part. These computer-based classification models can help in improving precision and reliability of analysis and reduce the probable misunderstandings, as well as making the diagnosis more effective. In this study, we tried to develop a robust model using Feature ensemble learning based on sparse autoencoders to classify healthy people and people with Parkinson disease (PWD) through proper representation of the vocal and speech datasets. The main aim was not only to perform a comparison between DNN and FESA-DNN but also to be benefitted from the best true accuracy of the proposed FESA-DNN model. Both FESA-DNN and DNN were tested for 10 runs of tenfold cross-validations and compared on the obtained averaged results. The results show that the proposed classifier outperforms the DNN on OPD dataset. Additionally, the proposed model is also comparable with the existing methods available in the literature. The experimental results and statistical analyses are pointed out that the proposed classifier is really useful and practical model for Parkinson's disease investigation. The future research will pay much attention to estimate accuracy of the proposed system in other medical classification problems.

References

1. Parkinson, J.: An essay on the shaking palsy. J. Neuropsychiatry Clin. Neurosci. **14**(2), 223–236 (2002). Published online: May 01
2. Langston, J.W.: Parkinson's disease: current and future challenges. Neurotoxicology, (4-5), pp. 443–450, 23 Oct 2002. https://doi.org/10.1016/S0161-813X(02)00098-0
3. Meissner, W.G., Frasier, M., Gasser, T., Goetz, C.G., Lozano, A., Paola Piccini, J.A., Obeso, O.R., Schapira, A., Voon, V., Weiner, David M., Tison, F., Bezard, E.: Priorities in Parkinson's disease research. Nature Rev. Drug Discov. **10**(377), 393 (2011)
4. Gallagher, D.A., Lees, A.J., Schrag, A.: What are the most important nonmotor symptoms in patients with Parkinson's disease and are we missing them? Mov. Disord. **25**(2493), 2500 (2010). https://doi.org/10.1002/mds.23394
5. American Parkinson Disease Association (APDA), Symptoms of Parkinson's disease. https://www.apdaparkinson.org/what-is-parkinsons/symptoms/ (2017). Accessed 21 Nov 2017
6. Mittel, C.S.: Parkinson's Disease, Overview and Current Abstracts, 1st edn. Nova Science Publishers. Inc., New York, USA (2003)
7. Schley, W.S., Fenton, E., Niimi, S.: Vocal symptoms in parkinson disease treated with levodopa: a case report. Ann. Otol. Rhinol. Laryngol. **91**(1), 119–121 (1982)
8. Little, M.A., McSharry, P.E., Hunter, E.J., Spielman, J., Ramig, L.O.: Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Trans. Biomed. Eng. **56**(4), 1015–1022 (2009). https://doi.org/10.1109/TBME.2008.2005954
9. Shahbaba, B., Neal, R.: Nonlinear models using dirichlet process mixtures. J. Mach. Learn. Res. **10**, 1829–1850 (2009)
10. Psorakis, I., Damoulas, T., Girolami, M.A.: Multiclass relevance vector machines: sparsity and accuracy. IEEE Trans. Neural Netw. **21**(10), 1588–1598 (2010)
11. Guo, P.-F., Bhattacharya, P., Kharma, N.: Advances in detecting Parkinsons disease. In: Zhang, D., Sonka, M. (eds.) Medical Biometrics. ICMB 2010. Lecture Notes in Computer Science, vol. 6165. Springer, Berlin, Heidelberg (2010)
12. Okan Sakar, C., Kursun, O.: Telediagnosis of parkinson's disease using measurements of dysphonia, J. Med. Syst. **34**(4), 591–599 (2010)
13. Das, R.: A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Syst. Appl. **37**, 1568–1572 (2010)
14. Luukka, P.: Feature selection using fuzzy entropy measures with similarity classifier. Expert Syst. Appl. **38**, 4600–4607 (2011). https://doi.org/10.1016/j.eswa.2010.09.133
15. Ozcift, A., Gulten, A.: Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Comput. Methods Programs Biomed. **104**(3), 443–451 (2011). https://doi.org/10.1016/j.cmpb.2011.03.018. Epub 30 Apr 2011
16. Li, D.-C., Liu, C.-W., Hu, S.C.: A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. Artif. Intell. Med. **52**, 45–52 (2011)
17. Åström, F., Koker, R.: A parallel neural network approach to prediction of Parkinson's Disease. Expert Syst. Appl. **38**(10), 12470–12474 (2011)

18. Spadoto, A.A., Guido, R.C., Carnevali, F.L., Pagnin, A.F., Falcao, A.X., Papa, J.P.: Improving Parkinson's disease identification through evolutionary-based feature selection. In: Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '11), pp. 7857–7860, Boston, Massachusetts, USA (2011)
19. Polat, K.: Classification of Parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering. Int. J. Syst. Sci. **43**(4), 597–609 (2012). https://doi.org/10.1080/00207721.2011.581395
20. Daliri, M.R.: Chi-square distance kernel of the gaits for the diagnosis of Parkinson's disease. Biomed. Signal Process. Control **8**, 66–70 (2013)
21. Chen, H.-L., Huang, C.-C., Yu, X.-G., Xu, X., Sun, X., Wang, G., Wang, S.-J.: An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. Expert Syst. Appl. **40**, 263–271 (2013)
22. Zuo, W.-L., Wang, Z.-Y., Liu, T., Chen, H.-L.: Effective detection of Parkinson's disease using an adaptive fuzzy k-nearest neighbor approach. Biomed. Signal Process. Control **8**(4), 364–373 (2013)
23. Zhang, Y.N.: Can a smartphone diagnose Parkinson disease? In: A Deep Neural Network Method and Telediagnosis System Implementation, Parkinsons Disease, vol. 2017, Article ID 6209703, 11 pp (2017). https://doi.org/10.1155/2017/6209703
24. Lu, Y., Zhang, L., Wang, B., Yang, J.: Feature ensemble learning based on sparse autoencoders for image classification. In: 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, pp. 1739–1745 (2014). https://doi.org/10.1109/IJCNN.2014.6889415
25. Lichman, M.: UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml (2013). Accessed 20 Nov 2017
26. Andrew, N.G.: CS294A Lecture Notes Sparse Autoencoder, p. 72 (2011). https://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf
27. Kuncheva, L.I.: Combining Pattern Classifiers, Methods and Algorithms, p. 544. Wiley (2004)
28. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 3, 226–239 (1998). https://doi.org/10.1109/34.667881
29. Little, M.A., McSharry, P.E., Roberts, S.J., Costello, D.A.E., Moroz, I.M.: Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Biomed. Eng. OnLine **6**, 23 (2007)

PCA Fusion for ANN-Based Diabetes Diagnostic

Sandeep Sangle, Pramod Kachare and Jitendra Sonawane

Abstract Diabetes is a result of inability to respond felicitously to insulin requirement for glucose regulation (sugar). In this paper, diabetes detection system is developed utilizing Principal Component Analysis (PCA) and Multilayer Perceptron Artificial Neural Network (MLPANN). Primary investigation focuses on combining source information and PCA transformed features under MLPANN framework. Confusion matrix based analysis has been performed to analysis the effect of source and PCA information fusion. In analysis standard UCI diabetes dataset, the maximum accuracy of 76.5% has been recorder for source features and accuracy of 85.2% with 6 level PCA features while fusion resulted in highest success rate of 87.8%. It acquires 15% and 3% relative accuracy increase when compared with source and PCA features used alone, respectively.

Keywords Principal component analysis · Artificial neural network · Confusion matrix · Feature fusion

1 Introduction

In 1965, World Health Organization (WHO) has laid down set of rules for diagnosis and classification of diabetes patients. Further in 1977, American Diabetes Association proposed fasting plasma Glucose tolerance test then practiced by WHO.

Human services experts frequently utilize most of these and other available diagnosis methods which requires human expertise to analyze and diagnosis diabetic record. Proposed system has following fundamental contributions:

- Dimensionality reduction using principal component analysis
- Classifier built using multilayer perception artificial neural network

S. Sangle (✉) · P. Kachare · J. Sonawane
Department of Electronics and Telecommunication Engineering,
Ramrao Adik Institute of Technology, Nerul, Navi Mumbai 400706, India
e-mail: sandeepsangle12@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_59

Design of diagnosis models for diabetes has been an active research area for the past decade. Most of the models found in literature are based on clustering algorithms and artificial neural networks (ANNs).

Kumari and Chitra are introduced diabetes diagnosis using support vector machine (SVM) and machine learning algorithms. This technique provides accuracy of 78% [1]. Ephzibah gives cost effect diabetes diagnosis model by combining Genetic (GA) and fuzzy logic best subset select and its improvement in accuracy. Standard machine learning dataset with 8 attributes for 769 cases was used to evaluate performance three best features selected using GA are explored using fuzzy classifier resulting in 87% diagnosis accuracy [2].

2 Proposed System

Most of literature uses ML for classification but no feature development. In this work have proposed classifier for principle features.

2.1 Principal Component Analysis

Principal Component Analysis (PCA) is the common name for a technique which elaborates fundamental mathematical concepts to transforms a number of possibly correlated variables into a smaller number of variables called Principal Components (PC). In general terms, PCA utilizes a vector space transform to decrease the dimensionality of huge data sets. Applying mathematical projection, the original data set, which may have involved number of variables, can be presented in just a few variables (the PC) [3].

PCA is the first choice of analysis to retrieve desired pattern in noisy environment another motivation behind PCA is possible reduction in feature dimensionality resulting in low dimension visualization and comparison. Mathematically PCA is data transformation using set of orthogonal basis vector called as eigenvector (orthogonal nature of transformation avoids redundant information) [4].

Step 1: Normalize the data: Initial step is to standardize the information that we have so that PCA works appropriately. This is finished by subtracting the particular means from the numbers in the separate segment. So the off chance that have two measurements X and Y, all X progress toward becoming x- and all Y move toward becoming y-. This delivers a dataset whose mean is zero [4].

Step 2: Calculate the covariance matrix: Since the dataset in this work is 2-dimensional, this will result in a 2 × 2 Covariance matrix [4], (Fig. 1).

$$\boldsymbol{Matrix}(\boldsymbol{Covariance}) = \left[\frac{\mathbf{Var}[\mathbf{X}1]}{\mathbf{Cov}[\mathbf{X}2, \mathbf{X}1]} \frac{\mathbf{Cov}[\mathbf{X}1, \mathbf{X}2]}{\mathbf{Var}[\mathbf{X}2]}\right] \tag{1}$$

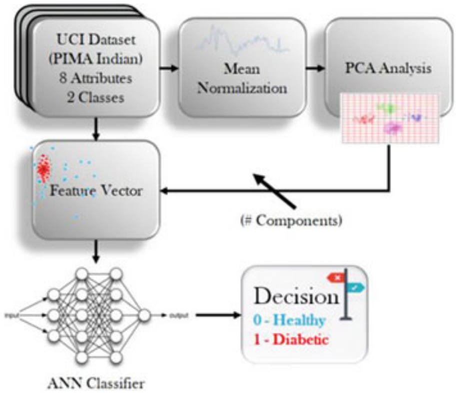


Fig. 1 The flow diagram of proposed system

Here,

Var[X1] = Cov[X1, X1] and Var[X2] = Cov[X2, X2].

Step 3: Calculate the eigenvalues and eigenvectors: From covariance matrix, next step is to compute the eigenvalues and eigenvectors. ƛ is an eigenvalue for a matrix A if the characteristic equation satisfied following condition [4],

$$\det(\lambda I - A) = 0 \tag{2}$$

where, I is the identity matrix of the same dimension as A which is a required condition for the matrix subtraction and 'det' is the determinant of the matrix [4].

For each eigenvalue ƛ, a corresponding eigenvector v, can be found by solving:

$$(\boldsymbol{\lambda I} - \boldsymbol{A})\,\boldsymbol{v} = \boldsymbol{0} \tag{3}$$

Step 4: Component selection and forming a feature vector: The eigenvalues are ordered from largest to smallest so that it gives us the significant components. Here comes the dimensionality reduction part. Hear a dataset have with n variables, and then we have the corresponding n eigenvalues and eigenvectors. It turns out that the eigenvector corresponding to the highest eigenvalue is the PC of the dataset and it is our decision to select appropriate no of PC to proceed with our analysis. So to reduce the dimensions, we select the first p eigenvalues and ignore the rest [4].

Step 5: Forming Principal Components: This is the final step where actually forms the PC which are extracted from feature vector [4].

$$\text{New Data} = (\text{Feature Vector})^{T} \times (\text{Scaled Data})^{T} \tag{4}$$

Here,

New Data is the Matrix consisting of the PC,

Feature Vector is the matrix we formed using the eigenvectors we chose to keep, and Scaled Data is the scaled version of original dataset.

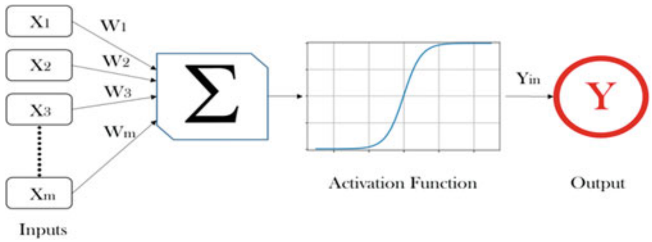


Fig. 2 Block diagram of artificial neural network system

2.2 Artificial Neural Network (ANN)

Neural systems are parallel processing gadgets, which essentially aims to model the brain. The fundamental target is to build a network that performs different computational assignments with reduced latency and higher accuracy than that of conventional network.

The proposed system implements ANN to explore Non linear abstraction of feature and class correlation. In particular, PCA features are trained and tested using Multilayer Perceptron (MLP) Back Propagation Neural Net (BPNN) with scaled conjugate gradient learning model. Many such models with various parameter settings were explored in search of the optimized classifier.

For the general model of ANN shown in Fig. 2, the net input can be calculated as follows:

$$\mathbf{Yin} = \sum \mathbf{x_i} \cdot \mathbf{w_i}, \quad \mathbf{i} = 1, 2, \ldots, \mathbf{m} \tag{5}$$

The output can be calculated by applying the activation function over the net input.

$$\mathrm{Y} = \mathbf{F(y\,in)} \tag{6}$$

$$\text{Output} = \text{function (net input calculated)}$$

ANN systems work through the enhanced weight values. The technique by which the improved weight esteems are accomplished is called learning. In the learning procedure attempt to instruct the system, how to deliver the outcomes when the relating input is displayed. When Learning is finished: the prepared neural system, with the refreshed ideal weights, need to have the capacity to create the output to exactness relating to an information design.

Table 1 Eight attributes available in pima-Indian dataset [5]

Sr. no.	Name of attributes
1	Number of time pregnancies
2	Glucose level
3	Blood pressure
4	Skin thickness
5	Insulin in body
6	BMI weight in kg
7	Diabetes pedigree function
8	Age (year)

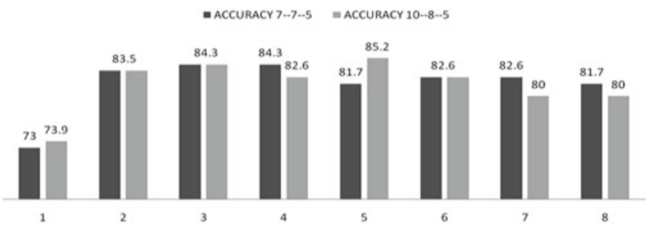


Fig. 3 Percentage accuracy using PCA + MLPANN for various no. of PCA

3 Results and Discussion

In this study, the PCA Features are classified using MLPANN for diagnosis of diabetes. The dataset is obtained from [5]. All patients in this dataset are Pima-Indian ladies. Binary classes in the dataset are represented using 8 clinical attributes as shown in Table 1. These attributes are identified as source features which identify a particular class of a sample.

Class index "1" implies diabetes patient and "0" represents non-diabetes or healthy record. It consists of clinical records of 768 participants which is further divided as 268 diabetic and 500 non-diabetic records, respectively.

Dataset consists of 768 samples out of which 230 are diabetic and 538 samples are healthy. 185-83 samples were used for training and testing of PCA–ANN Diagnosis system for diabetes samples, respectively. 353–147 samples were used for training and testing of PCA–ANN Diagnosis system for healthy samples. The dimensionally compressed features (PCA features) are applied at the input of MLPANN. The different dimensionalities have been analyzed for optimum compression level. Figures 3 and 4 shows the percentage accuracy for various PC.

As indicating percentage accuracy in Fig. 4, Architecture 7-7-5 provides overall better accuracy than 10-8-5 considering PCA with source features except one PCA Component. PCA component three five, seven and eight for Architecture 7-7-5 and 10-8-5 provides exact same accuracy respectively 83.5%, 84.3%, 86.1% and 83.5%.

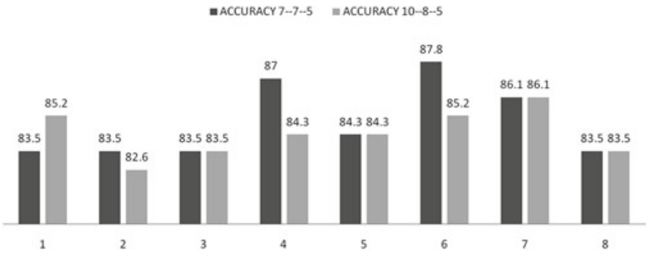


Fig. 4 Percentage accuracy using **Source data + PCA + MLPANN** for various No. of PCA components

Table 2 The obtain value of sensitivity and specificity by using PCA and ANN diagnosis system of diabetes

Method	Sensitivity in %	Specificity %
LS-SVM [6]	73.91	80
GDA-LS-SVM [6]	79.16	83.33
LDA-ANFIS [7]	83.33	85.18
LDA-MWSVM [8]	83.33	93.75
Source + PCA – ANN (Proposed)	83.33	89.41

For six PCA component analyses Architecture 7-7-5 is achieved overall highest accuracy of 87.8%. Architecture 10-8-5 provides highest accuracy for seven PCA components that is 86.1%. In summary, PCA along with source features are provides better accuracy as compare to PCA features. If we combine of PCA with source features accuracy increased up to 87.8% (Table 2).

The proposed algorithm is compared with state of the art diabetes diagnosis systems. The comparative studies of Sensitivity and specificity is shown in Table 3, also The Confusion matrix based comparative analysis is shown in Table 4. In this study, we have analyzed several MLPANN architectures to evaluate optimum configuration for current problem. Percentage accuracy for various numbers of PCA component using two MLPANN architectures shown in the Figs. 3 and 4. In that Fig. 3 utilizes only PCA features but Fig. 4 is utilizes source features with PCA.

As presented in Fig. 3, for PCA component two, three six and eight for Architecture 7-7-5 and 10-8-5 provides exact same accuracy respectively 83.5%, 84.3%, 82.6% and 81.7%. For architecture 10-8-5 in five PCA component analyses is achieved overall highest accuracy of 85.2%. Architecture 7-7-5 provided highest accuracy of 84.3% for 3 and 4 PCA Component.

In this comparative studies we are analysis four another methods. By k. Polat and Güne [6] introduce two methods first one is least square support vector machine (LS-SVM) gives accuracy, sensitivity and specificity are 78.21%, 73.91%, and 80% respectively [9]. And second method is Linear discriminate analysis (LDA-LS-SVM) it gives 79.16%, 49.16%, and 83.33% respectively [6].

Table 3 The obtain value of confusion matrix using PCA and ANN

Sr. No	Output/desired	Number of healthy	Number of dieses	Method
1	Number of healthy	25	9	Source + PCA –ANN (Proposed)
	Number of dieses	5	76	
2	Number of healthy	45	5	LDA-MWSVM [8]
	Number of dieses	3	25	
3	Number of healthy	46	4	LDA-ANFIS [7]
	Number of dieses	8	20	
4	Number of healthy	45	5	GDA-LS-SVM [6]
	Number of dieses	3	25	
5	Number of healthy	44	6	LS-SVM [6]
	Number of dieses	11	17	

Calisir and Dongantekin [8] done experiment using LDA and Adaptive Neuro-Fuzzy Interference System (LDA-ANFIS) he gives accuracy 89.74% and sensitivity is 83.33% specificity is 93.75%. All above methods are used 90-10% partition are used for training and testing in diabetes diagnosis system [8].

4 Conclusion

This paper focuses on accuracy improvement using source information with PCA transformation fusion for diabetes diagnosis. Different PCA dimensionalities have been analyzed for optimum features order. PCA compressed features alone provides diagnosis accuracy of 85.2% for five-layer (8I-10 N-8 N-5L-2O) bottleneck architecture MLPANN classifier. On the other hand, additions of source features to PCA compressed features provide accuracy as high as 87.8%. A relative increase of 3% has been observed by addition of source information. Finally, an optimum diabetes diagnosis system with source information fused with 6 level PCA features gives maximum accuracy for 5 layer (14I-7 N-7 N-5L-2O) MLPANN classifier. Its performance can be further improved by indentifying and incorporate various other parameters and by increasing training data.

Acknowledgements Author(s) are thankful to ***Kaggle and UCI machine learning repository*** for using their publically available dataset in the experimentation work stated in this paper.

References

1. Anuja Kumari, V., Chitra, R.: Classification of diabetes disease using support vector machine. **3**(2), 1797–1801 (2013)
2. Ephzibah, E.P.: Cost effective approach on feature selection using genetic algorithms and fuzzy logic for diabetes diagnosis. Int. J. Soft Comput. **2**(1) (2011)
3. Richardson, M.: Principal component analysis. http://people.maths.ox.ac.uk/richardsonm/SignalProcPCA.pdf Accessed 3 May 2013. Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik@ ntf. uni-lj. si 6 (2009): 16
4. Lindasy, L.S.: A tutorial on principal components analysis. http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf. Accessed 21 Nov 2017
5. Blake, C.L., Merz, C.J.: UJI repository of machine learning dataset (1996). http://www.ics.uci.edu/~mlearn/MLRepository.html
6. Polat, K., Güne, S.: Principles component analysis, fuzzy weighting pre-processing and artificial immune recognition system based diagnostic system for diagnosis of lung cancer. Expert Syst. Appl. **34**(1), 214–221 (2008)
7. Dogantekin, E., et al.: An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS. Digit. Signal Process. **20**(4), 1248–1255 (2010)
8. Çalisir, D., Dogantekin, E.: An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier. Expert Syst. Appl. **38**(7), 8311–8315 (2011)
9. Polat, K., Gunes, S.: An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digit. Signal Process. **17**(4), 702–710 (2007)

MHD Flow with Heat and Mass Transfer Over a Radiating Cone Due to a Point Sink in Presence of Partial and Solutal Slips

Nasreen Bano Shaikh, B. B. Singh and S. R. Sayyed

Abstract An analysis has been carried out to study the effects of velocity, thermal and solutal slips on a magnetohydrodynamic, steady, and incompressible laminar boundary layer flow with heat and mass transfer over a radiating cone due to a point sink. The problem has been solved by using a semi-analytical method called DTM-Padé. Graphical representations are obtained for velocity, temperature, and concentration distributions for various values of the governing parameters like suction/injection parameter s, magnetic parameter M, velocity slip $\mathscr{L}$, thermal slip δ_1, concentration slip δ_2, radiation parameter R, Prandtl number Pr, and Schimdt number Sc. Also, the numerical results obtained for skin-friction coefficient have been compared with the corresponding results available in the literature and a good conformity has been found in between them.

1 Introduction

The heat and mass transfer analysis of fluid flow under the influence of magnetic field, and velocity, thermal, and solutal slips in presence of radiation effect has been the subject of considerable interest due to its various geophysical, geothermal, and engineering applications. A boundary layer slip flow problem arises in polishing of artificial heart valves and internal cavities. Because of this reason only, the authors [1–6] studied the effects of these governing parameters by taking into consideration of various physical conditions. Also, representative studies dealing with the

N. B. Shaikh · B. B. Singh (✉) · S. R. Sayyed
Department of Mathematics, Dr. Babasaheb Ambedkar Technological University, Lonere, Raigad 402103, Maharashtra, India
e-mail: bbsingh@dbatu.ac.in

N. B. Shaikh
e-mail: snasreenbano@yahoo.in

S. R. Sayyed
e-mail: srsayyed786@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_60

effects of momentum, thermal and solutal slips can be found in the works of various researchers like [7–14]. Further, some pertinent radiative heat transfer studies for different configurations have been reported by researchers [15–19].

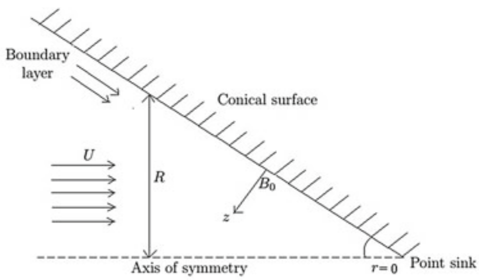
The objective of the present paper is to deal with the problem of MHD flow and radiative heat and mass transfer over a cone due to point sink in presence of partial and solutal slips by using the combination of differential transform method (DTM) developed by Zhou [20] with the Padé approximation. This technique is very effective as it is free from the process of discritization, linearization, and perturbation unlike the regular Taylor series method. On account of this fact only, DTM has been used by numerous authors [21–24] to deal with different problems for varying geometrical configurations.

In the present study, the numerical results obtained in respect of skin-friction coefficient have been compared with the results of Guled and Singh [6] who tackled the same problem in the absence of slip conditions by using homotopy analysis method (HAM). The results have been found in precise agreement. The effects of various governing parameters on velocity, temperature, and concentration profiles have also been analyzed graphically.

2 Problem Statement and Mathematical Formulation

Let us consider a steady, laminar, and axisymmetric flow of an electrically conducting incompressible fluid inside a circular cone at rest with the hole at the vertex. The boundary layer flow is due to the presence of the hole which is regarded as a three dimensional point sink (see Fig. 1). In the flow process, the magnetic Reynolds number is assumed to be very small, so as to ignore the effects of induced magnetic field as compared to the applied magnetic field. A magnetic field B_0 is applied in the z-direction which is fixed relative to the fluid. It is also assumed that the injected fluid and the boundary layer fluid possess the same physical properties. The wall temperature (T_w) and concentration (C_w), and likewise, the ambient temperature (T_∞) and

Fig. 1 Schematic diagram of the flow model



concentration (C_∞) are also assumed to be constant. Under the above assumptions, the governing similarity boundary layer equations for flow, heat, and mass transfer are (cf. [6]):

$$\begin{aligned} f''' - ff'' + 4(1 - f'^2) + M(1 - f') = 0, \\ \left(1 + \frac{4}{3}R\right)\theta'' - Prf\theta' = 0, \\ \phi'' - Scf\phi' = 0. \end{aligned} \tag{1}$$

under the boundary conditions

$$\begin{aligned} f(0) = s, \quad f'(0) = \mathscr{L} f''(0), \quad \theta(0) = 1 + \delta_1\theta'(0), \quad \phi(0) = 1 + \delta_2\phi'(0), \\ f'(\infty) = 1, \quad \theta(\infty) = 0, \quad \phi(\infty) = 0, \end{aligned} \tag{2}$$

where f' is velocity profile; θ is temperature profile; ϕ is concentration profile; $M = \frac{2\sigma B_0^2 r^3}{m_1\rho}$ is the local magnetic parameter; $R = \frac{4\gamma^* T_\infty^3}{k_1 k^*}$ is the radiation parameter; $Pr = \frac{\nu}{\alpha}$ is the Prandtl number; $Sc = \frac{\nu}{D}$ is the Schmidt number; $\mathscr{L} = L\frac{m_1^{1/2}}{(2\nu r^3)^{1/2}}$ is the velocity slip parameter; $\delta_1 = K_1\frac{m_1^{1/2}}{(2\nu r^3)^{1/2}}$ is the thermal slip parameter; $\delta_2 = K_2\frac{m_1^{1/2}}{(2\nu r^3)^{1/2}}$ is the solutal slip parameter; $s = V_w\left(\frac{2r^3}{m_1\nu}\right)^{1/2}$ is the local mass transfer parameter, where $s < 0$ corresponds to suction and $s > 0$ corresponds to injection, and the prime denotes derivative with respect to η.

The Eq. (1) along with boundary conditions (2) are highly nonlinear similarity boundary layer equations to be solved by DTM-Padé treatment in the present paper.

3 DTM-Padé Treatment

An arbitrary function $u(x)$ analytic in the domain T, can be written in the Taylor series form about the point $x = x_0$, as

$$u(x) = \sum_{k=0}^{\infty} \frac{1}{k!}\left[\frac{d^k u(x)}{dx^k}\right]_{x=x_0} (x - x_0)^k. \tag{3}$$

The differential transform of $u(x)$ is defined as

$$U(k) = \frac{1}{k!}\left[\frac{d^k u(x)}{dx^k}\right]_{x=x_0}, \tag{4}$$

where $u(x)$ is the original function and $U(k)$ is the transformed function.

4 Application

Taking the one-dimensional differential transform of each term of Eq. (1) with the help of Table 1, the following recurrence relations are obtained:

$$F(k+3) = \frac{1}{(k+1)(k+2)(k+3)} \left\{ \sum_{r=0}^{k} \Big[(r+1)(r+2)F(r+2)F(k-r) + 4(r+1) \right.$$
$$\left. (k-r+1)F(r+1)F(k-r+1) \Big] + M(k+1)F(k+1) - \delta(k)(4+M) \right\}, \tag{5}$$

$$G(k+2) = \frac{3Pr}{(3+4R)(k+1)(k+2)} \sum_{r=0}^{k} (k-r+1)F(r)G(k-r+1), \tag{6}$$

and

$$H(k+2) = \frac{Sc}{(k+1)(k+2)} \sum_{r=0}^{k} (k-r+1)F(r)H(k-r+1). \tag{7}$$

We can consider the boundary conditions in Eq. (2) as follows:

$$f(0) = s, \quad f'(0) = 1 + \mathscr{L} f''(0), \quad f''(0) = 2\alpha_1,$$
$$\theta(0) = 1 + \delta_1\theta'(0), \quad \theta'(0) = \alpha_2, \quad \phi(0) = 1 + \delta_2\phi'(0), \quad \phi'(0) = \alpha_3,$$

so that the differential transforms of these boundary conditions are

$$F(0) = s, \quad F(1) = 1 + 2\mathscr{L}\alpha_1, \quad F(2) = \alpha_1, \quad G(0) = 1 + \delta_1\alpha_2, \tag{8}$$
$$G(1) = \alpha_2, \quad H(0) = 1 + \delta_2\alpha_3, \quad H(1) = \alpha_3.$$

Table 1 The operations for the one-dimensional differential transform method

Original function	Transformed function
$u(x) = u_1(x) \pm u_2(x)$	$U(k) = U_1(k) \pm U_2(k)$
$u(x) = \lambda u_1(x)$	$U(k) = \lambda U_1(k)$, λ is a constant
$u(x) = \frac{du_1(x)}{dx}$	$U(k) = (k+1)U_1(k+1)$
$u(x) = \frac{d^r u_1(x)}{dx^r}$	$U(k) = (k+1)(k+2)\cdots(k+r)U_1(k+r)$
$u(x) = u_1(x)u_2(x)$	$U(k) = \sum_{r=0}^{k} U_1(r)U_2(k-r)$
$u(x) = \frac{du_1(x)}{dx}\frac{du_2(x)}{dx}$	$U(k) = \sum_{r=0}^{k}(r+1)(k-r+1)U_1(r+1)U_2(k-r+1)$
$u(x) = u_1(x)\frac{du_2(x)}{dx}$	$U(k) = \sum_{r=0}^{k}(k-r+1)U_1(r)U_2(k-r+1)$
$u(x) = u_1(x)\frac{d^2u_2(x)}{dx^2}$	$U(k) = \sum_{r=0}^{k}(k-r+2)(k-r+1)U_1(r)U_2(k-r+2)$

From the relations (5), (6), (7) and boundary conditions (8), we recursively calculate all the $F(k)$'s, $G(k)$'s, and $H(k)$'s. Using all the $F(k)$'s, $G(k)$'s, and $H(k)$'s, we can get the series solutions for f, θ and ϕ.

5 Results and Discussion

By using the DTM-Padé technique, the coupled nonlinear Eq. (1) have been solved with the help of the boundary conditions (2). To validate the accuracy of the present method, the values of skin-friction coefficient ($f''(0)$) have been compared with those reported by Guled and Singh [6] in Table 2, who tackled the same problem in the absence of partial slips and solutal slip by using homotopy analysis method (HAM). The results have been found in precise agreement. Further, the effects of various governing parameters like radiation parameter (R), velocity slip parameter ($\mathscr{L}$), thermal slip parameter (δ_1), and solutal slip parameter (δ_2) on velocity, temperature, and concentration profiles are studied graphically in Figs. 2, 3, 4, 5, 6, 7, 8.

From Table 2, it is obsereved that $f''(0)$ increases with the increasing values of suction parameter ($s < 0$) and magnetic parameter (M), but decreases with increasing values of injection parameter ($s > 0$). This is because the momentum boundary layer thickness gets reduced with the enhancement of suction and magnetic parameter, thereby increasing the skin friction. But, an opposite trend is observed for injection.

The velocity profiles $f'(\eta)$ for various values of velocity slip parameter ($\mathscr{L}$) are sketched in Figs. 2 and 3 for injection and suction, respectively. In both the cases of suction and injection, it is seen that with the increasing values of $\mathscr{L}$, the fluid velocity increases monotonically. As the slip parameter increases, more fluid slips near the wall and the flow gets accelerated, thereby increasing the fluid velocity near the wall and hence decreasing the boundary layer thickness. On the other hand, for distances away from the wall, an opposite behavior is seen.

The effects of thermal radiation (R) on temperature profiles are shown in Fig. 4. It is found that with an increase in radiation parameter (R), the temperature increases

Table 2 Comparability of skin-friction coefficient ($f''(0)$) by DTM-Pade[15/15] with Guled and Singh [6] by HAM and Takhar et al. [4] for $\mathscr{L} = \delta_1 = \delta_2 = R = 0$

M	s	Present	Guled and Singh [6]	Takhar et al. [4]
0.5	0	2.3797	2.3827	2.392
	1	1.96066	1.9117	1.973
	2	1.54922	1.5232	1.5529
0.5	−2	3.54054	3.6172	3.6162
	−1	3.00699	2.9554	3.0231

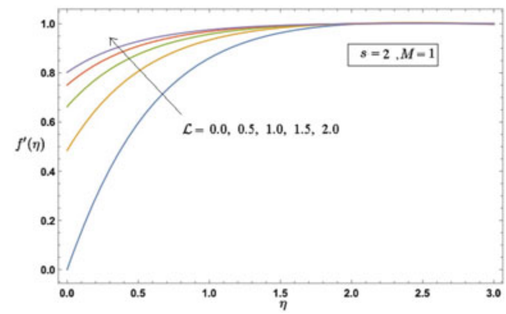


Fig. 2 Variation of velocity profile w.r.t. $\mathscr{L}$ in case of injection

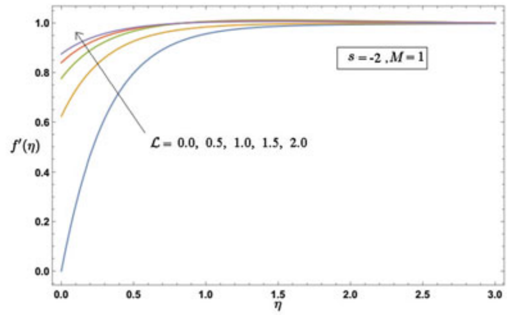


Fig. 3 Variation of velocity profile w.r.t. $\mathscr{L}$ in case of suction

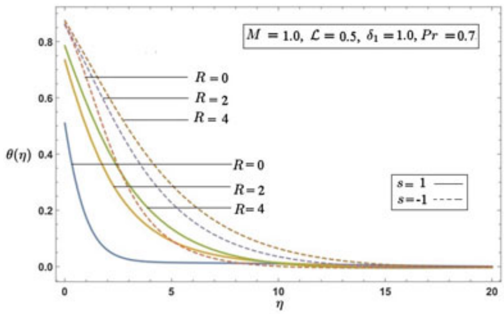


Fig. 4 Variation of temperature profile w.r.t. R in case suction/injection

in both the cases of suction and injection. This is due to the fact that the thermal boundary layer thickness increases with increasing R.

Figure 5 represents the temperature profiles for varying velocity slip parameter ($\mathscr{L}$) in both the cases of suction and injection. It is observed that temperature increases with the increase in slip parameter. An increase in slip parameter generates friction force and allows more fluid to slip past the wall. Due to existence of the friction force, temperature increases.

The effect of thermal slip parameter (δ_1) on temperature profile for the cases of suction/injection is displayed in Fig. 6. The figure shows a decrease in temperature

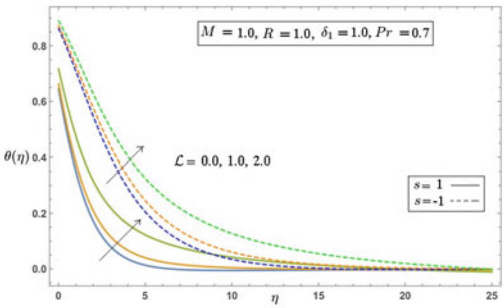


Fig. 5 Variation of temperature profile w.r.t. $\mathscr{L}$ in case of suction/injection

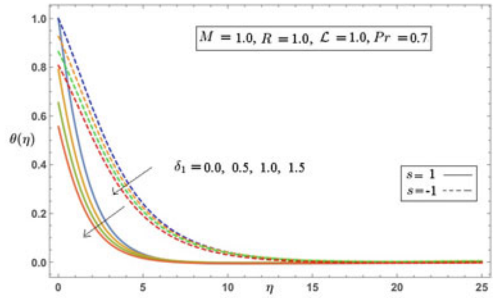


Fig. 6 Variation of temperature profile w.r.t. δ_1 in case of suction/injection

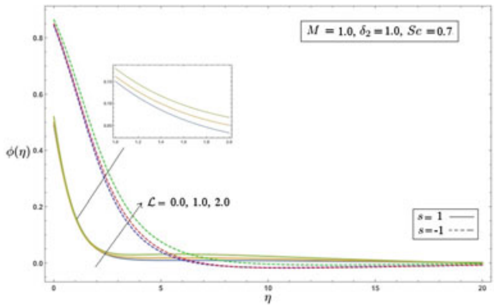


Fig. 7 Variation of concentration profile w.r.t. $\mathscr{L}$ in case of suction/injection

with an increase in thermal slip. This is because with the increase of thermal slip, less heat is transferred from the wall to the fluid.

The effect of momentum slip parameter ($\mathscr{L}$) on the concentration profile for both suction and injection is illustrated in Fig. 7. From the figure, it is obvious that the concentration gets enhanced with an increase in $\mathscr{L}$.

Finally, a weak reduction in concentration is observed from Fig. 8 for a significant rise in solutal slip parameter δ_2.

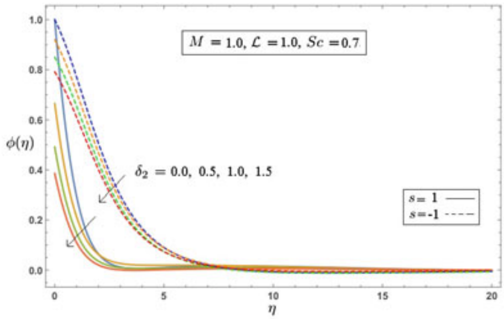


Fig. 8 Variation of concentration profile w.r.t. δ_2 in case of suction/injection

6 Concluding Remarks

The summary of the significant results is given below:

1. The skin-friction coefficient $f''(0)$ increases with magnetic parameter (M) and suction ($s < 0$), but it decreases with injection ($s > 0$).
2. The velocity profile exhibits an increasing trend with momentum slip parameter ($\mathscr{L}$) for both suction and injection.
3. The temperature profile enhances with the increase in the radiation parameter (R) and momentum slip parameter ($\mathscr{L}$). But, the thermal slip parameter (δ_1) reduces the temperature profile for both suction and injection.
4. The concentration profile shows an increase with momentum slip parameter ($\mathscr{L}$) whereas it shows a decrease with solutal slip parameter (δ_2).

References

1. Choi, S.H., Wilhelm, H.E.: Self-similar magnetohydrodynamic diffuser flows with induced magnetic fields. Phys. Fluids **20**(11), 1821–1826 (1977)
2. Rosenhead, L.: Laminar Boundary Layers. Clarendon Press, Oxford, UK (1963)
3. Takhar, H.S., Surma Devi, C.D., Nath, G.: MHD flow with heat and mass transfer due to point sink. Indian J. Pure Appl. Math. **17**(10), 1242–1247 (1986)
4. Eswara, A.T., Roy, S., Nath, G.: Unsteady MHD forced flow due to a point sink. Acta Mech. **145**(1–4), 159–172 (2000)
5. Zhang, J., Fang, T., Zhong, Y.: Analytical solution of magnetohyrodynamic sink flow. Appl. Math. Mech. **32**, 1221 (2011)
6. Guled, C.N., Singh, B.B.: Homotopy analysis of the radiation effect on MHD flow with heat and mass transfer due to a point sink. Adv. Math. Phys. Article ID 7036728, 14 (2016)
7. Andersson, H.I.: Slip flow past a stretching surface. Acta Mech. **158**, 121–125 (2002)
8. Ariel, P.D., Hayat, T., Asghar, S.: The flow of an elastico-viscous fluid past a stretching sheet with partial slip. Acta Mech. **187**, 29–35 (2006)
9. Abbas, Z., Wang, Y., Hayat, T., Oberlack, M.: Slip effects and heat transfer analysis in a viscous fluid over an oscillatory stretching sheet. Int. J. Numer. Methods Fluids **59**, 443–458 (2009)

10. Mukhopadhyay, S., Gorla, R.S.R.: Effects of partial slip on boundary layer flow past a permeable exponential stretching sheet in presence of thermal radiation. Heat Mass Transf. **48**, 1773–1781 (2012)
11. Turkyilmazoglu, M.: Exact multiple solutions for the slip flow and heat transfer in a converging channel. J. Heat Transf. **137**, 101301–1 (2015)
12. Ibrahim, W., Shankar, B.: MHD boundary layer flow and heat transfer of a nanofluid past a permeable stretching sheet with velocity, thermal and solutal slip boundary conditions. Comput. Fluids **75**, 1–10 (2013)
13. Laxmi, T.V., Shankar, B.: Radiative boundary layer flow and heat transfer of nanofluid over a nonlinear stretching sheet with slip conditions and suction. Jordan J. Mech. Ind. Eng. **10**(4), 285–297 (2016)
14. Uddin, M.J., Bg, O.A., Uddin, M.N.: Multiple slip and variable transport property effects on magneto-hydromagnetic dissipative thermo-solutal convection in porous media. ASCE J. Aerosp. Eng. **29**(5) (2016). https://doi.org/10.1061/(ASCE)AS.1943-5525.0000614
15. Sparrow, E.M., Cess, R.D.: Radiation Heat Transfer. Brooks Cole Publishing Company, Belmont, California (1970)
16. Vafai, K., Hadim, H.A., Howell, J.R.: Radiative transfer in porous media. In: Handbook of Porous Media. CRC Press, New York (2000)
17. Sedeek, M.A., Salem, A.M.: Laminar mixed convection adjacent to vertical continuously stretching sheet with variable viscosity and variable thermal diffusivity. Heat Mass Transf. **41**(12), 1048–1055 (2005)
18. Mukhopadhyay, S.: Effects of radiation and variable fluid viscosity on flow and heat transfer along a symmetric wedge. J. Appl. Fluid Mech. **2**(2), 29–34 (2009)
19. Vyas, P., Rai, A.: Radiative variable fluid properties flow due to a point sink inside a cone filled with porous medium. Appl. Math. Sci. **6**(87), 4307–4317 (2012)
20. Zhou, J.K.: Differential Transformation and Its Applications for Electric Circuits. Huazhong University Press, Wuhan, China (1986)
21. Chen, C.K., Ho, S.H.: Solving partial differential equations by two dimensional differential transform method. Appl. Math. Comput. **106**, 171–179 (1999)
22. Ayaz, F.: Solution of the system of differential equations by differential transform method. Appl. Math. Comput. **147**(2), 547–567 (2004)
23. Darania, P., Ebadian, A.: A method for the numerical solution of the integro-differential equations. Appl. Math. Comput. **188**(1), 657–668 (2007)
24. Rashidi, M.M.: The modified differential transform method for solving MHD boundary layer equations. Comput. Phys. Commun. **180**, 2210–2217 (2009)

MHD Stagnation-Point Dissipative Flow in a Porous Medium with Joule Heating and Second-Order Slip

S. R. Sayyed, B. B. Singh and Nasreen Bano

Abstract The present paper deals with the MHD stagnation-point dissipative flow in a porous medium over a flat plate with variable wall temperature. The effects of viscous dissipation, Joule heating, and second-order slip on the flow field have been studied both numerically and graphically for several values of governing parameters. The physical model of the problem is governed by coupled partial differential equations reducible to a set of coupled nonlinear ordinary differential equations (ODEs) using similarity transformations. The system of the coupled nonlinear ODEs has been solved analytically using optimal homotopy analysis method (OHAM). The results obtained in the present analysis have been compared with the results available in the literature, and have been found in excellent agreement.

1 Introduction

The study of the boundary-layer flow and heat transfer towards a stagnation point has attracted the attention of several researchers due to its various natural and industrial applications such as flows over the tips of submarines, front tip of rockets and aircrafts, etc. On account of this reason only, the most pioneering work in this context has been carried out by Hiemenz [1], which was further extended by Eckert [2], Beard and Walters [3], and many others.

The magneto-hydrodynamic (MHD) stagnation-point flows with heat transfer problems pertaining to porous media have their numerous thermal engineering applications such as geothermal energy recovery, crude oil extraction, thermal insulation,

S. R. Sayyed · B. B. Singh (✉) · N. Bano
Department of Mathematics, Dr. Babasaheb Ambedkar Technological University, Lonere, Raigad 402103, Maharashtra, India
e-mail: bbsingh@dbatu.ac.in

S. R. Sayyed
e-mail: srsayyed786@gmail.com

N. Bano
e-mail: snasreenbano@yahoo.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_61

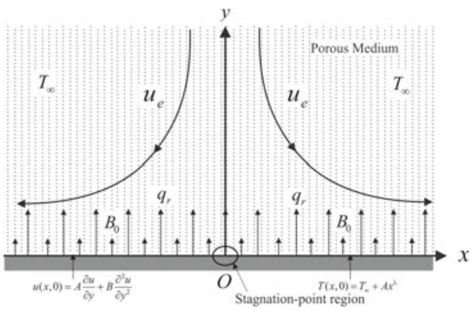


Fig. 1 Physical model and coordinate system

thermal energy storage, food processing, solar power collectors, etc. As a consequence of this, only extensive reviews relating to this topic have been provided in book [4]. Further extension of the work in the context of porous media has also been carried out by various researchers such as Yih [5], Raptis and Takhar [6], Kechil and Hashim [7], Bhatti et al. [8], Kudenatti et al. [9], etc.

In all the studies mentioned above have been carried out by neglecting viscous dissipation, Joule heating effect, and no-slip boundary conditions on the surface of the plate have been applied. But the effect of viscous dissipation, Joule heating with second-order slip plays an important role in engineering. On account of this fact, only the researcher [10–13] have done pioneering work in this flow field by taking various geometrical configurations.

The objective of the present paper is, therefore, to investigate the two-dimensional steady laminar forced MHD stagnation-point flow on a flat plate with variable wall temperature in a porous medium; together with viscous dissipation, Joule heating and second-order slip model as proposed by Wu [11]. The problem has been analyzed by using the optimal homotopy analysis method (OHAM) [14].

2 Problem Statement and Mathematical Formulation

Following the Yih [5] model for the porous medium and by introducing the boundary-layer approximation, the governing continuity, momentum and energy equations can be written as follows (see Fig. 1):

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \tag{1}$$

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = \nu\frac{\partial^2 u}{\partial y^2} + u_e\frac{du_e}{dx} - \frac{\sigma B_0^2}{\rho}(u - u_e) - \frac{\nu}{K}(u - u_e). \tag{2}$$

$$u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y} = \alpha\frac{\partial^2 T}{\partial y^2} + \frac{\nu}{C_p}\left(\frac{\partial u}{\partial y}\right)^2 + \frac{\sigma B_0^2}{\rho C_p}(u_e - u)^2. \tag{3}$$

The boundary conditions are defined as follows:

$$\begin{array}{llll} y = 0; & v = v_w, & u = u_{slip}, & T = T_w = T_\infty + Ax^\lambda, \\ y \to \infty; & u = u_e = ax, & T = T_\infty, & \end{array}. \quad (4)$$

The symbols used in the abovementioned equations are as follows:
u, v: velocity components along $x-$ and $y-$axes, respectively; P: pressure; ρ: fluid density; $\nu(=\frac{\mu}{\rho})$: kinematic viscosity, μ: coefficient of fluid viscosity; K: porosity parameter; σ: electrical conductivity; B_0: applied magnetic field along $y-$axis; T: temperature of the fluid and the porous medium which are in local thermal equilibrium; $\alpha(=\frac{k}{\rho c_p})$: equivalent thermal diffusivity, k: coefficient of thermal conductivity; c_p:specific heat at constant pressure; v_w is the uniform surface mass flux positive (i.e., $v_w > 0$) for blowing and negative (i.e., $v_w < 0$) for suction; $u_{slip} = A\frac{\partial u}{\partial y} + B\frac{\partial^2 u}{\partial y^2}$: velocity slip, A, B are constant; λ; wall temperature exponent, and $u_e(= ax)$: free stream velocity, a: positive constant.

The stream function satisfying Eq. (1) is defined as $u = \frac{\partial \psi}{\partial y}$ and $v = -\frac{\partial \psi}{\partial x}$.

The following similarity variables have been defined:

$$\begin{array}{l} \eta = y\sqrt{\left(\frac{a}{\nu}\right)}, \quad \psi = \sqrt{(a\nu)}xf(\eta), \quad \theta = \dfrac{T - T_\infty}{T_w - T_\infty}, \\ u = axf'(\eta) \quad \text{and} \quad v = -\sqrt{a\alpha}f(\eta), \end{array} \quad (5)$$

where η is similarity variable and primes denote differentiation with respect to η.

The following ordinary differential equations can be obtained by substituting (5) into Eqs. (2) and (3):

$$f''' + ff'' + (1 - f'^2) + (M^2 + \Omega)(1 - f') = 0, \quad (6)$$

$$\frac{1}{Pr}\theta'' + f\theta' - \lambda f\theta' + Ec\left[(f'')^2 + M^2(1 - f')^2\right] = 0, \quad (7)$$

with the boundary conditions which are transformed to

$$\begin{array}{lll} f(0) = s, & f'(0) = \beta f''(0) + \gamma f'''(0), & \theta(0) = 1; \\ f'(\infty) = 1, & \theta(\infty) = 0, & \end{array} \quad (8)$$

where primes denote differentiation with respect to η.

The other parameters governing the problem, e.g., Pr, Ω, M, s, β, γ, Ec and λ are defined below:
$Pr = \frac{\nu}{\alpha}$ (Prandtl number), $\Omega = \frac{\nu}{Ka}$ (permeability parameter), $M = \sqrt{\frac{\sigma B_0^2}{\rho a}}$ (magnetic parameter), $s = -\frac{v_w}{\sqrt{a\nu}}$ (mass transfer parameter), $\beta = A\sqrt{(\frac{a}{\nu})} > 0$ (dimensionless first-order velocity slip parameter), $\gamma = B\left(\frac{a}{\nu}\right) < 0$ (dimensionless second-order velocity slip parameter), and $Ec = \frac{u_e^2(x)}{C_p[T_w(x) - T_\infty]}$ (Eckert number). Here, $s > 0$ for suction, $s < 0$ for blowing and $s = 0$ represents the impermeable surface.

3 Analytic Solution by OHAM

The series solutions of Eqs. (6)–(7) with boundary conditions (8) are constructed by employing optimal homotopy analysis method (OHAM). The suitable initial guesses, auxiliary functions and linear operators are selected as follows:
The initial guesses:

$$f_0(\eta) = s + \eta + \left(\frac{1}{1+\beta-\gamma}\right)(e^{-\eta} - 1); \qquad \theta_0(\eta) = e^{-\eta}, \tag{9}$$

the auxiliary functions:

$$H_f(\eta) = e^{-\eta}; \qquad H_\theta(\eta) = e^{-\eta}. \tag{10}$$

and the linear operators:

$$\mathscr{L}_1(f) = \frac{d^3 f}{d\eta^3} + \frac{d^2 f}{d\eta^2}; \qquad \mathscr{L}_2(\theta) = \frac{d^2\theta}{d\eta^2} - \theta, \tag{11}$$

which have the following properties:

$$\mathscr{L}_1(c_1 + c_2\eta + c_3 e^{-\eta}) = 0; \qquad \mathscr{L}_2(c_4 e^{\eta} + c_5 e^{-\eta}) = 0, \tag{12}$$

where $c_1,\ c_2,\ c_3,\ c_4$, and c_5 are arbitrary constants.

The analytic approximations of coupled nonlinear ODEs (6) and (7) satisfying the boundary condition (8) can be obtained using BVPh 2.0 with the aid of the auxiliary operators (11), the initial guesses (9), and the auxiliary functions (10).

4 Optimal Convergence Control Parameters

The series solution by HAM comprises nonzero convergence control parameters c_0^f and c_0^θ which determine the convergence region. To achieve the optimal values of c_0^f and c_0^θ, the concept of minimization of averaged squared residual has been used (see [14]).

The averaged squared residual of the governing equations are defined by

Table 1 Individual averaged squared errors using $c_0^f = -0.198367$ and $c_0^\theta = -0.73158$ where $\Omega = M = 2,\ s = -1, \beta = 0.2, \gamma = -0.2, Pr = 7, Ec = 0.05$ and $\lambda = 1$

m (order of approximation)	10	20	30	40
$\mathscr{E}_m^f$	8.90113×10^{-6}	2.3204×10^{-8}	3.16664×10^{-10}	3.97481×10^{-11}
$\mathscr{E}_m^\theta$	1.64953×10^{-4}	7.61×10^{-5}	4.632×10^{-6}	3.69509×10^{-6}

$$\mathscr{E}_m^f(c_0^f, c_0^\theta) = \frac{1}{(k+1)} \sum_{j=0}^{k} \left[\mathscr{N}_f \left(\sum_{i=0}^{m} f_i \right) \Bigg|_{\eta = j\delta\eta} \right], \tag{13}$$

$$\mathscr{E}_m^\theta(c_0^f, c_0^\theta) = \frac{1}{(k+1)} \sum_{j=0}^{k} \left[\mathscr{N}_\theta \left(\sum_{i=0}^{m} f_i, \sum_{i=0}^{m} \theta_i \right) \Bigg|_{\eta = j\delta\eta} \right], \tag{14}$$

where k is an integer and $\mathscr{N}_f$ and $\mathscr{N}_\theta$ are nonlinear operators corresponding to Eqs. (6) and (7).

Here, the optimal values of c_0^f and c_0^θ are determined by minimizing the total error $\mathscr{E}_m^t(c_0^f, c_0^\theta)$ as suggested by [14], where

$$\mathscr{E}_m^t(c_0^f, c_0^\theta) = \mathscr{E}_m^f(c_0^f, c_0^\theta) + \mathscr{E}_m^\theta(c_0^f, c_0^\theta). \tag{15}$$

This error has been calculated by using the command "**GetOptiVar**"of BVPh 2.0.

A case has been considered, wherein $\Omega = M = 2,\ s = -1, \beta = 0.2, \gamma = -0.2,$ $Pr = 7,\ Ec = 0.05$, and $\lambda = 1$. The corresponding optimal values for sixth order of approximations are $c_0^f = -0.201485$ and $c_0^\theta = -2.79858$, and the total averaged squared residual is $\mathscr{E}_6^t(c_0^f, c_0^\theta) = 8.12084 \times 10^{-4}$.

Table 1 shows the individual averaged squared residual for these optimal values of convergence control parameters c_0^f and c_0^θ for various orders of approximation. It is found that the individual averaged squared residual error decreases for higher orders of approximation.

5 Results and Discussion

Tables 2 and 3 have been prepared to validate the present results with results previously existing in published literature. It is evident from these tables that the present values of local skin friction coefficient $f''(0)$ and local Nusselt number at $Pr = 1$, $\Omega = 0, \beta = 0, \gamma = 0$, and for different values of M and s are in excellent agreement with those reported by Bhatti et al. [8], Kechil and Hashim [7] and Yih [5].

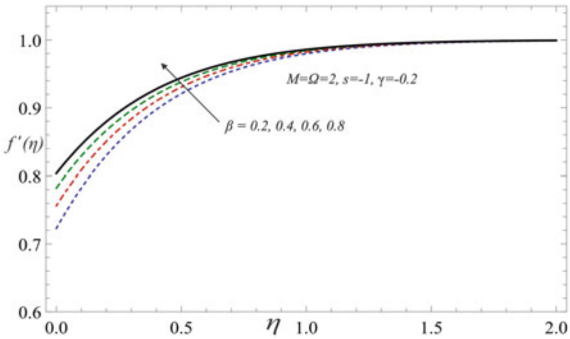
Table 2 Comparative values of $f''(0)$ for various values of M at $Pr = 1$, $\Omega = 0$, $\beta = 0$, $\gamma = 0$

s	M	Present results	Bhatti et al. [8]	Kechil and Hashim [7]	Yih [5]
−1	0	0.756575	0.7565	0.7559794662	0.756575
	1	1.116421	1.1164	1.1164292350	1.116421
	2	1.877621	1.8776	1.8776221492	1.877620
	5	4.667525	–	4.6675255187	4.667525
	10	9.585914	–	9.5859131993	9.585913

Table 3 Comparison of $-\theta'(0)$ for various values of M at $Pr = 1$ and $\Omega = R = \beta = \gamma = 0$

λ	s	M	Present results	Kechil and Hashim [7]	Yih [5]
0	−1	0	0.11675	0.11677	0.116752
		1	0.1400	0.14000	0.140002
		2	0.17312	0.17312	0.173124

Fig. 2 Velocity profile for various values of β



In order to analyze the effects of first-order velocity slip parameter (β), second-order velocity slip parameter (γ), and Eckert number (Ec) on flow and heat transfer, the Figs. 2, 3, 4, 5, and 6 have been plotted.

Figures 2, 3, 4, and 5 display the influence of velocity slip on the velocity and temperature profiles. It is observed that velocity profile is an increasing function and the temperature profile is a decreasing function of first-order slip parameter. On the other hand, an opposite trend can be seen in case of second-order velocity slip parameter.

Figure 6 depicts the effects of Eckert number (Ec) on the temperature distribution. It is perceived that both the temperature distribution and the wall temperature increases as viscous dissipation parameter increases. This phenomenon happens because of the fact that the liquid elements store more energy for large values of Ec. So, the consideration of viscous dissipation in energy equation may play an important role in gaining the temperature.

Fig. 3 Temperature profile for various values of β

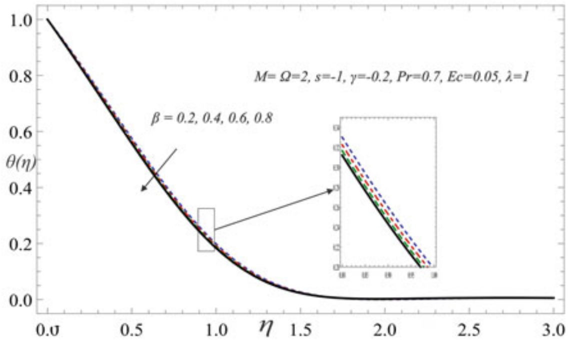


Fig. 4 Velocity profile for various values of γ

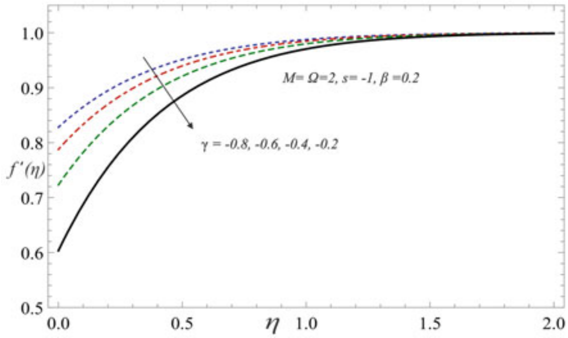
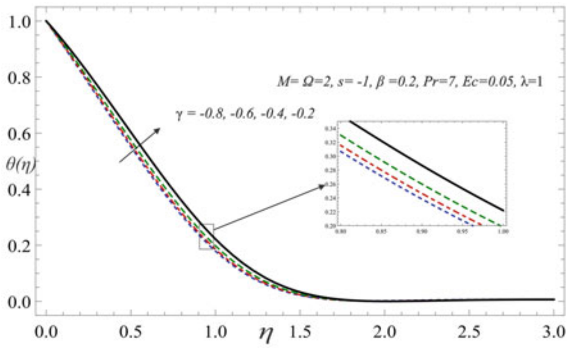


Fig. 5 Temperature profile for various values of γ



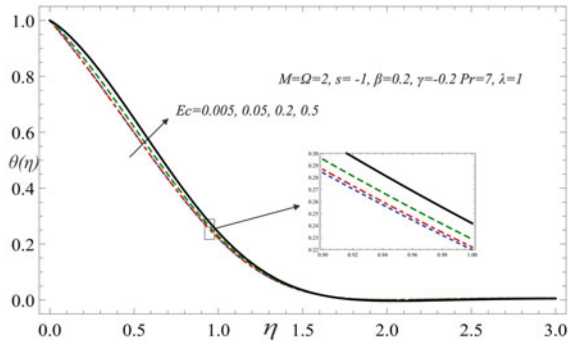


Fig. 6 Temperature profile for various values of Ec

6 Concluding Remarks

The significant results are summarized below:

1. There occurs an increase in the fluid flow along with an increase in the values of the β.
2. There occurs a decrease in the velocity along with the increasing values of γ.
3. An increasing trend is observed in the temperature profiles along with the increasing values of the parameters Ec and γ.
4. With an increase in the values of the parameters M and β, there occurs an increase in the heat transfer rate.

References

1. Hiemenz, K.: Die Grenszchicht an einem in den gleichförmingen Flussigkeitsstorm eingetauchen geraden Kreiszylinder. Dinglers Polytech J. **326**, 321–324 (1911)
2. Eckert, E.R.G.: Die berechnung des wärmeübergangs in der laminaren grenzschicht umströmter körper VDI Forschungsheft, Berlin (1942)
3. Beard, D.W., Waltersm, K.: Elastico-viscous boundary-layer flows. I. Two dimensional flow near a stagnation point. Math. Proc. Cambr. Philos. Soc. **60**(3), 667–674 (1964). https://doi.org/10.1017/s0305004100038147
4. Nield, D.A., Bejan, A.: Convection in Porous Media. Springer, New York (2006)
5. Yih, K.A.: The effect of uniform suction/blowing on heat transfer of magnetohydrodynamic Hiemenz flow through porous media. Acta Mech. **130**(3–4), 147–158 (1998)
6. Raptis, A.A., Takhar, H.S.: Flow through a porous medium. Mech. Res. Commun. **14**(5–6), 327–329 (1987). https://doi.org/10.1016/0093-6413(87)90049-8
7. Kechil, S.A., Hashim, I.: Approximate analytical solution for MHD stagnation-point flow in porous media. Commun. Nonlinear Sci. Numer. Simul. **14**(4), 1346–1354 (2009). https://doi.org/10.1016/j.cnsns.2008.02.007
8. Bhatti, M.M., Abbas, T., Rashidi, M.M.: A new numerical simulation of MHD stagnation-point flow over a permeable stretching/shrinking sheet in porous media with heat transfer. Iran. J. Sci. Technol. Trans. A: Sci. 1–7 (2016). https://doi.org/10.1007/s40995-016-0027-6

9. Kudenatti, R.B., Kirsur, S.R., Nargund, A.L., Bujurke, N.M.: Similarity solutions of the MHD boundary layer flow past a constant wedge within porous media. Math. Probl. Eng. **2017**, 1–11 (2017). https://doi.org/10.1155/2017/1428137
10. Khan, M.I., Hayat, T., Khan, M.I., Alsaedi, A.: A modified homogeneous-heterogeneous reactions for MHD stagnation flow with viscous dissipation and Joule heating. Int. J. Heat Mass Transf. **113**, 310–317 (2017)
11. Wu, L.: A slip model for rarefied gas flows at arbitrary Knudsen number. Appl. Phys. Lett. **93**, 253103 (2008). https://doi.org/10.1063/1.3052923
12. Fang, T., Yao, S., Zhang, J., Aziz, A.: Viscous flow over a shrinking sheet with a second order slip flow model. Commun. Nonlinear Sci. Numer. Simul. **15**, 1831–1842 (2010). https://doi.org/10.1016/j.cnsns.2009.07.017
13. Rahman, M.M.: Effects of second-order slip and magnetic field on mixed convection stagnation-point flow of a Maxwellian fluid: multiple solutions. J. Heat Transf. **138**, 122503-1 (2016). https://doi.org/10.1115/1.4034161
14. Zhao, Y., Liao, S.: HAM-based package BVPh 2.0 for nonlinear boundary value problems. In: Liao, S. (ed.) Advances in Homotopy Analysis Method. World Scientific Press (2013)

Design Optimization of 10 nm Channel Length InGaAs Vertical Gate-All-Around Transistor (Nanowire)

Shreyas Kulkarni, Sangeeta Joshi, Dattatray Bade and Subha Subramaniam

Abstract This paper proposes a cylindrical vertical Gate-All-Around Transistor with nanowire of compound III-V semiconductor material $In_{0.53}Ga_{0.47}As$ n-type device with channel length of 10 nm. The effect of variation of channel diameter and spacer length on the performance of the device is simulated. The device gives an acceptable Subthreshold Slope and Drain Induced Barrier Lowering along with satisfactory I_{ON}/I_{OFF} ratio. The device is simulated in Sentaurus Synopsys using Hydrodynamic model for III-V semiconductors with Poisson equation to give the transfer characteristics.

Keywords Vertical Gate-All-Around (VGAA) · InGaAs · Nanowire · Synopsys Sentaurus TCAD

1 Introduction

The first transistor was invented by John Bardeen, Walter Brattain and William Shockley in 1956, since then the electronics went into an unprecedented era. This revolution was further enhanced later with the first integrated circuits (ICs) fabrication in 1960s. According to Gordon E. Moores prediction in 1965 the number of transistors per chip roughly doubled every 2 years by downscaling the transistor size [1]. The Moore's law became a major goal for semiconductor industry for higher speed and

S. Kulkarni (✉) · S. Joshi · D. Bade
Vidyalankar Institute of Technology, Mumbai, India
e-mail: shreyas.v.kulkarni@ieee.org

S. Joshi
e-mail: sangeeta.joshi@vit.edu.in

D. Bade
e-mail: dattatray.bade@vit.edu.in

S. Subramaniam
Shah and Anchor Kutchhi Engineering College, Mumbai, India
e-mail: subha.sakec@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_62

power efficiency by increasing the packing density by using downscaling of the transistors.

Although downscaling improves IC technology, as the MOSFET gate length enters the nanometer regime, the Short Channel Effects (SCE) like threshold voltage (V_t) rolloff and Drain Induced Barrier Lowering (DIBL) becomes significant [2]. This limits the scaling of the planar Silicon on Insulator (SOI) MOSFETs. Also due to the relatively low carrier mobility of the silicon, the MOSFET device performance degrades. To overcome these limitations various novel materials and device structures like carbon nanotubes [3] silicon nanowire transistor [4], new channel materials like pure silicon-germanium, III-V semiconductors are being explored. Amongst all the novel technology in More-than-Moore, the nanowire-based transistor gives high performance which have been fabricated [5].

The use of FinFET's over traditional planar structure gives a better performance and scalalability but as the technology approaches the sub-10 nm regime where the channel length is around 10 nm and below, the use of FinFET structure limits the performance of the device. Here the short channel effects due to reduced fin width is prominent. To overcome this issue the most ideal form of the multigate structure which is Gate-All-Around (GAA) Transistor are being explored.

Nanowire to form channel gives a better electrostatic control as compared to FinFET's. The Vertical GAA MOSFET's using the silicon nanowire has been demonstrated where the Subthreshold slope is 95 mV/dec and Drain Induced Barrier Lowering is 25 mV/V [6]. But still the limitation of silicon with its low mobility is prominent in these sub-10 nm devices. To overcome these issues the III-V semiconductor materials are being researched to replace the silicon technology. The use of Silicon-Germanium also is a viable option as it gives a higher mobility than its pure silicon counterpart. The use of conical shaped nanowire structure as well as different gate materials has been used to achieves a subthreshold slope of 69.3 mV/dec and a high drive current [7]. Vertical Nanowire Transistors have been demonstrated to give a low Drain Induces Barrier Lowering with a significant I_{ON}/I_{OFF} ratio. This is achieved by using compound III-V semiconductor material In0.53 Ga0.47 As.

The use of III-V semiconductor materials is widely recognized as an alternative for traditional material. These materials are considered due to their high mobilities than strained silicon in ITRS 2.0 [8]. As the silicon technology cannot be utilised for channel length of 10 nm and below, the use of III-V materials and Ge is needed.

In this paper Vertical Gate-All-Around transistor with $In_{0.53}Ga_{0.47}As$ as a channel of length of 10 nm is proposed. Here the channel diameter and spacer length are optimised to give satisfactory performance in terms of subthreshold slope and Drain Induced Barrier Lowering (DIBL). This device is simulated in Synopsys Sentaurus TCAD software. The details of the structure and simulation is explained in Sect. 2. The results and discussion is given in Sect. 3 and conclusion is given in last section.

2 Device Structure and Simulation of Gate-All-Around Transistor

The device is simulated by using Poisson equation and quantum potential equation for electrons and holes in Sentaurus TCAD. The simulated result gives output in terms of electrostatic potential in various regions depending on the voltage applied. The device is simulated for channel length of 10 nm to get the transfer characteristics and extract values of Subthreshold Slope (SS), Drain Induced Barrier Lowering (DIBL), I_{ON} and I_{OFF}.

The device has $In_{0.53}Ga_{0.47}As$ as the channel material with InP as the substrate with resistivity of 7×10^2 Ω cm. The oxide layer over substrate of InP is HfO_2 with k = 25. The contacts are of Au and the material used for spacer between the contacts is of HfO2. This device is Vertical in structure which reduces the footprint of the device drastically. The Si doped region has a doping of 5 × 1019 /cm^3. The detailed values of the device is given in Table 1. The Fig. 1 gives the 3-D view of the device which is simulated. Figure 2 gives cross-sectional cut of the Vertical Gate-All-Around MOSFET. The various parts of the device are labelled.

The device is simulated in Synopsys Sentaurus TCAD v.2014. The device is simulated by using Poisson equation and to accomodate the quantum effects, the Density Gradient Equations for electrons and holes are used as eQuantumPotential and hQuantumPotential in Sentaurus Device. This solves Schrdinger carrier quantization model and quantum potential equations self consistently with Poisson equation. The Shockley-Hall-Reed recombination is also calculated in simulation for the feasibility of the device. The device is optimised for Channel Diameter and Spacer Length. The device is simulated in Sentaurus Device where the transfer characteristics, subthreshold slope, Drain Induced Barrier Lowering (DIBL) values are extracted. The mole fraction of InAs used is 0.53 which is the most stable state where there is low lattice mismatch with InP when nanowire is formed.

Table 1 Device dimensions

Parameters	Value
Channel length	10 nm
Channel diameter	6 nm
Spacer length	10 nm
Height of the contacts	10 nm
Oxide height	10 nm
Gate oxide	2 nm
N++ Silicon doped layer	2 nm
Doping concentration	5×10^{19} cm^{-3}
Length of the substrate	90 nm
Height of the substrate	20 nm
Width of the substrate	90 nm

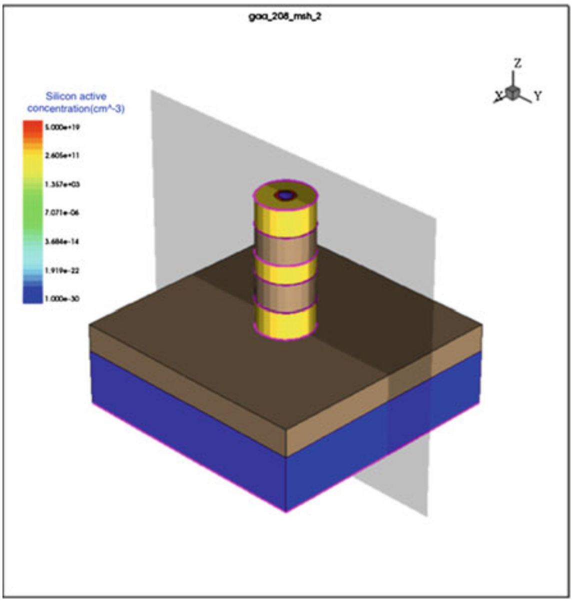


Fig. 1 Vertical Gate-All-Around transistor

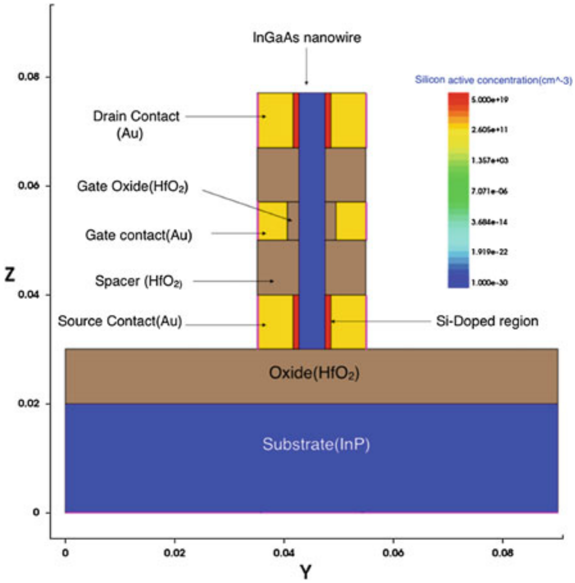


Fig. 2 Cross-sectional view of transistor

The device is simulated for vd (linear) of 0.05 V and vd (saturation) of 1 V with vgs of 1 V. The device is first simulated for the vdlin thereafter vdsat for finding the Drain Induced Barrier Lowering of the device for a given dimension. Here the device is first optimised for diameter of channel then for spacer length, i.e., length between source-gate contact and gate-drain contact. The output of the simulation gives the transfer characteristics, i.e., Id versus Vg graph and the SS and DIBL are extracted from the graph. The device dimensions are given in Table 1.

3 Results and Discussion

The Diameter Optimization of the device is done for channel length of 10 nm. To understand the impact of variation of diameter on performance of the device, we have chosen diameter of 6 nm, diameter of 5 nm and diameter of 4 nm to simulate the device.

The results extracted from the change in diameter for the channel length of 10 nm is given in the Fig. 3. The effect of reducing the diameter decreases the drive current of the device with the leakage current variation being small. The drive current of the device for 6 nm diameter is 1×10^{-5} A whereas for the diameter of 4 nm and diameter of 5 nm we achieve drive current of 8×10^{-6} A and 5×10^{-6} A respectively. The leakage current for the diameter variation i.e. for 4 nm diameter, 5 nm diameter and 6 nm diameter is 2.8×10^{-9} A, 2.5×10^{-9} A and 2×10^{-9} A respectively. The diameter of 6 nm gives better drive current and leakage current.

The Drain Induced Barrier Lowering (DIBL) occurs in small channel length MOSFET's where the dimension of source/drain contact or the channel doping is too low. This leads to electrostatic interaction between the source and drain where the punchthough leakage and loss of gate control occurs. The lowest value of DIBL extracted is 30 mV/V for diameter of 6 nm whereas the diameter of 4 nm gives 44 mV/V DIBL (Fig. 4 and Tables 2 and 3).

With the diameter of 6 nm, the Spacer length optimization is carried out as the capacitance between the source-gate contact and drain-gate contact changes with change in the spacer length. The impact of the change in capacitance is reflected in

Table 2 Drive current and leakage current values for channel length of 10 nm for different diameters

Diameter (nm)	I_{ON} (A)	I_{OFF} (A)
4	5×10^{-6}	2×10^{-9}
5	8×10^{-6}	2.5×10^{-9}
6	1×10^{-5}	2.8×10^{-9}

Table 3 Diameter Optimization

Diameter (nm)	SS (mV/dec)	DIBL (mV/V)	I_{ON}/I_{OFF}
4	86.079	44	10^3
5	82.838	–	10^3
6	81.413	30	10^4

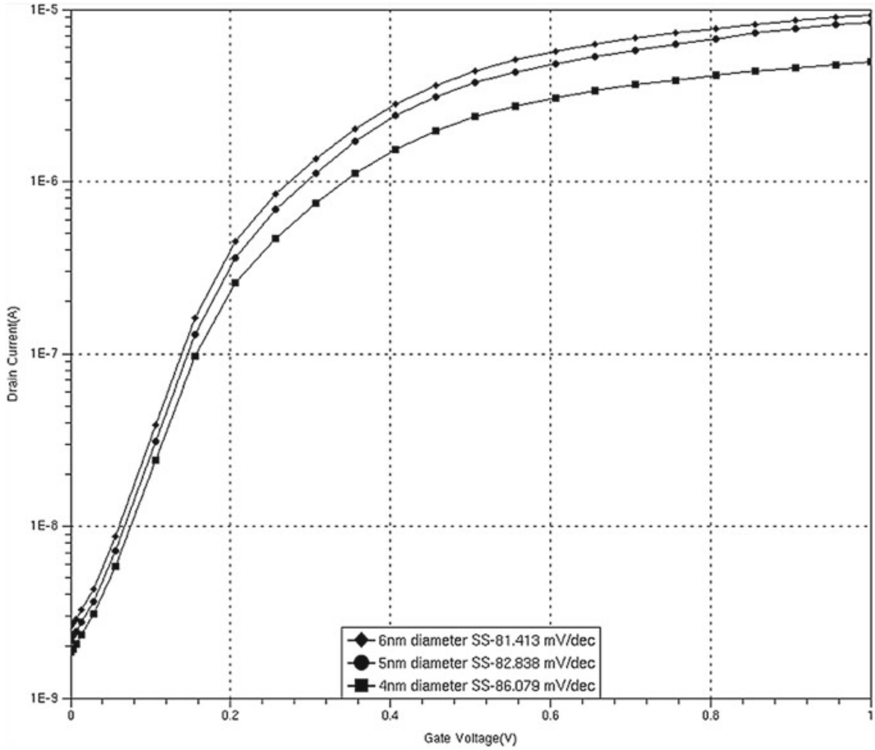


Fig. 3 Effect of diameter for channel length 10 nm with spacer length of 10 nm with a contact height of 10 nm

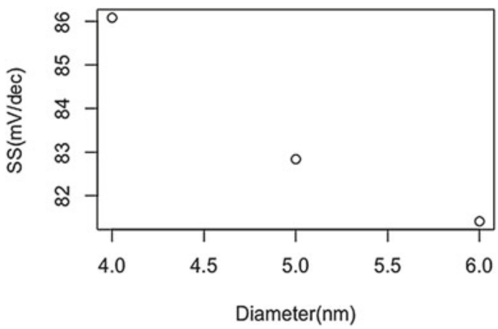


Fig. 4 Effect of diameter on subthreshold slope for channel length of 10 nm

transfer characteristics of the device. The Fig. 5 gives the I_d/V_g graph for the spacer length of 7, 8, 9 and 10 nm. Here the drive current for the spacer length of 7 nm is higher but the subthreshold slope is poor as compared to the spacer length of 10 nm. The graph of SS versus spacer length is given in Fig. 6 (Table 4).

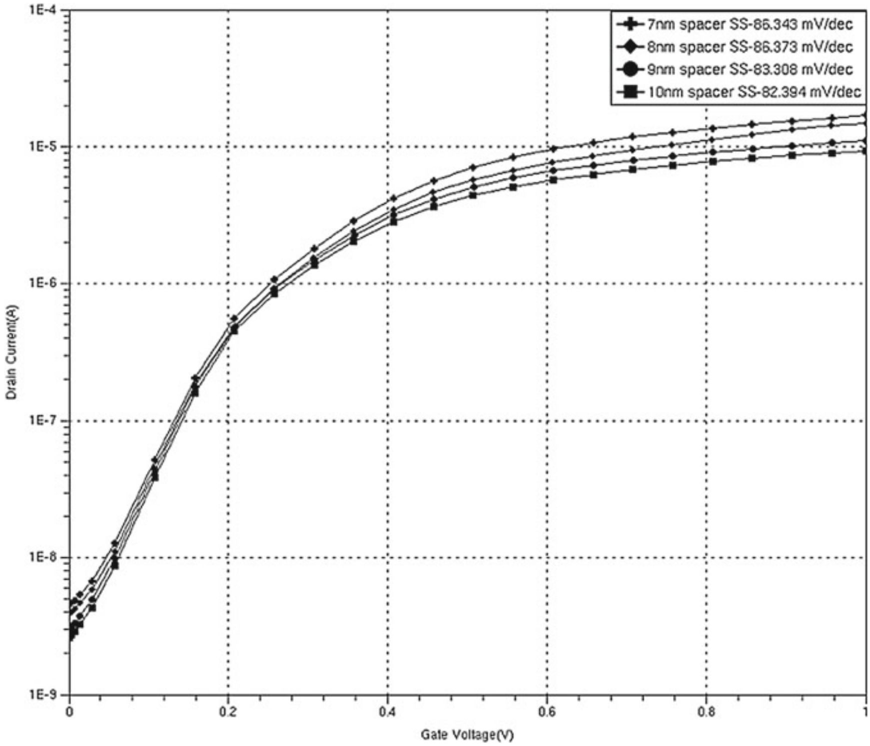


Fig. 5 Effect of spacer lengths with channel length 10 nm and diameter of 6 nm with a contact height of 10 nm

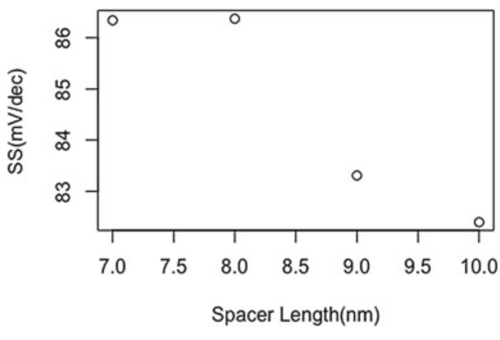


Fig. 6 Subthreshold slope variation with space length for channel length of 10 nm

The lowest subthreshold slope achieved is for spacer length of 10 nm is 82.394 mV/dec. Although the drive current achieved is relatively low of 1×10^{-5} A while considering drive current for spacer length of 7 nm which is 2×10^{-6} A, the value of subthreshold slope is much higher for spacer length of 7 nm. The value

Table 4 Spacer Optimisation

Diameter (nm)	SS (mV/dec)	DIBL (mV/V)	I_{ON}/I_{OFF}
10	82.394	28	10^4
9	83.308	31	10^4
8	86.373	–	10^4
7	86.343	25	10^4

Table 5 Drive current and leakage current values for channel length of 10 nm for different spacer lengths

Diameter (nm)	I_{ON} (A)	I_{OFF} (A)
10	1×10^{-5}	2.8×10^{-9}
9	0.3×10^{-5}	3×10^{-9}
8	1.5×10^{-5}	4×10^{-9}
7	2×10^{-5}	4.8×10^{-9}

of DIBL is also low for spacer length of 10 nm so this is selected as the optimised length. The values of drive current and leakage current is given in Table 5. Hence the channel diameter of 6 nm and spacer length of 10 nm gives the best performance of Vertical Gate-All-Around MOSFET.

4 Conclusion

We have simulated a Vertical Gate-All-Around Transistor device structure of 10 nm channel length on Sentaurus Synopsys, where the $In_{0.53}Ga_{0.47}As$ nanowire is used as a channel material. This device gives best performance for diameter of 6 nm and spacer length of 10 nm, where the subthreshold slope of 81 mV/dec, DIBL of 30 mV/dec, I_{ON} and I_{OFF} of 1×10^{-5} A and 2.8×10^{-9} A respectively is achieved. This device can be used for post CMOS technology.

Acknowledgements The authors would like to thank Indian Nanoelectronics Users Program (INUP), IIT-B for accepting our project and helping us in the same.

References

1. Moore, G.E.: Progress in digital integrated electronics. In: Electron Devices Meeting, vol. 21, pp. 11–13 (1975)
2. Taur, Y., Ning, T.H.: Fundamentals of Modern VLSI Devices. Cambridge University Press (2013)
3. Guo, J., Lundstrom, M., Datta, S.: Performance projections for ballistic carbon nanotube field-effect transistors. Appl. Phys. Lett. **80**(17), 3192–3194 (2002)

4. Wang, J., Polizzi, E., Lundstrom, M.: A computational study of ballistic silicon nanowire transistors. In: IEEE International Electron Devices Meeting, 2003. IEDM'03 Technical Digest, pp. 29–5. IEEE (2003)
5. Cui, Y., Zhong, Z., Wang, D., Wang, W.U., Lieber, C.M.: High performance silicon nanowire field effect transistors. Nano Lett. **3**(2), 149–152 (2003)
6. Guerfi, Y., Larrieu, G.: Vertical silicon nanowire field effect transistors with nanoscale gate-all-around. Nanoscale Res. Lett. **11**(1), 210 (2016)
7. Jena, B., et al.: Conical surrounding gate MOSFET: a possibility in gate-all-around family. Adv. Nat. Sci.: Nanosci. Nanotechnol. **7**(1), 015009 (2016)
8. ITRS 2.0 Executive Report 2015 Edition
9. Becker, J.A., Shive, J.N.: The transistor—a new semiconductor amplifier. Electr. Eng. **68**(3), 215–221 (1949)
10. Doyle, B.S., Datta, S., Doczy, M., Hareland, S., Jin, B., Kavalieros, J., Chau, R.: High performance fully-depleted tri-gate CMOS transistors. IEEE Electron Device Lett. **24**(4), 263–265 (2003)
11. Riel, H., Wernersson, L.-E., Hong, M., del Alamo, J.A.: III-V compound semiconductor transistors from planar to nanowire structures. MRS Bull. **39**(08), 668–677 (2014). Materials Research Society, Aug 2014
12. Subramaniam, Subha, Joshi, Sangeeta M., Awale, R.N.: Analytical modeling and numerical simulation of novel double-gate InGaAs vertical nanowire transistor device for threshold voltage tuning and improved performance. Eng. Sci. Technol. Int. J. **19**(4), 1857–1863 (2016)
13. Moore, G.E.: Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp. 114. IEEE Solid-State Circuits Soc. Newsl. **20**(3), 33–35 (2006)
14. Del Alamo, J.A., et al.: Nanometer-scale III-V MOSFETs. IEEE J. Electron Devices Soc. **4**(5), 205–214 (2016)
15. Lin, J., et al.: Record maximum transconductance of 3.45 mS/m for III-V FETs. IEEE Electron Device Lett. **37**(4), 381–384 (2016)
16. Larrieu, G., Guerfi, Y., Han, X.L., Clment, N.: Sub-15 nm gate-all-around field effect transistors on vertical silicon nanowires. Solid-State Electron. (2017)
17. Zhao, et. al.: Vertical nanowire InGaAs MOSFETs fabricated by a top-down approach. In: IEDM 2013
18. Zhou, X., Li, Q., Tang, C.W., Lau, K.M.: Inverted-type InGaAs metal-oxide-semiconductor high-electron-mobility transistor on Si substrate with maximum drain current exceeding 2 A/mm. Appl. Phys. Express **5**(10), 104201 (2012)
19. Sentaurus Device User Guide, Synopsys (2011)
20. Sentaurus Structure Editor User Guide, Synopsys (2011)

Design of Micro-heater on 3D-SnO_2 Gas Sensor

Gajendrasingh Y. Rajput, Manoj S. Gofane and Sandip Dhobale

Abstract Design of the heater on resistive gas sensors plays an important role since the performance of the gas sensor depends on temperature of the sensing materials. Heater on the SnO_2 gas sensor is designed in 3D geometry. The meander structure of heater is designed in such a way that the distribution of temperature is uniform on the sensor. COMSOL Multiphysics 5.0 simulating tool based on finite element method is used to study Joules heating in heater. Temperature of the sensor is maintained in the range of 617–621 K (344–348 °C). Uniform distribution of temperature is found on the surface of the sensors with variation of ± 2 °C.

Keywords SnO_2 gas sensor · Micro-heater · COMSOL Multiphysics

1 Introduction

SnO_2 is widely used in chemical sensing applications such as gas sensing and glucose sensing since it has enormous properties such as highly active surface formed due to dual valance nature of cations (Sn^{+2}/Sn^{+4}). Dual valance nature creates oxygen vacancies on surface and hence forming unsaturated covalent bonds (dangling bonds). These bonds are highly unstable and therefore easily transfer electrons with chemical species adsorbed on the surface of SnO_2 [1–3]. Several metal oxides have been reported as gas sensors with different operating temperatures with various gas entities.

G. Y. Rajput (✉) · M. S. Gofane · S. Dhobale
Department of Electronics Engineering, Ramrao Adik Institute of Technology, Nerul, Navi Mumbai, India
e-mail: gyrajput@gmail.com

M. S. Gofane
e-mail: manojgfn92@gmail.com

S. Dhobale
e-mail: dhobale.sandip@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_63

Table 1 Operating temperature of various metal oxide gas sensors

Sr. no.	Metal Oxide	Operating temperature
1	Tin Oxide (SnO_2)	300–400 °C
2	Cerium Oxide (CeO_2)	400–600 °C
3	Zinc Oxide (ZnO)	300–500 °C
4	Vanadium Oxide (V_2O_5)	300–400 °C
5	Titanium Oxide (TiO_2)	200–300 °C
6	Tungsten Oxide (WO_3)	500–550 °C
7	Indium Oxide (In_2O_3)	300–400 °C

Table 1 gives details of the operating temperatures of a few well-reported metal oxides. SnO_2 is a most favourable material reported yet for gas sensing applications and has shown better results [4–7].

Design of micro-heater plays a very important role in gas sensors and actuators, as the temperature uniformity in the sensor is a very crucial factor for the sensitivity and stability of the sensor. Thermal stability of the micro-heater depends much on parameters of the micro-heater such as properties of the heating material and geometry of heater. Uniform distribution of temperature of sensing material is provided by micro-heaters, and hence it is essential to design micro-heater with specific geometry and specific heating materials. Several attempts have been done to design gas sensors using COMSOL Multiphysics with different geometries and different materials for constructing heaters on the gas sensors. Velmathi et. al and Selvakumar et. al have designed only micro-heaters with various 2D geometries such as single and double meander, fan shaped and square structure and S-shaped structure with different distributions of temperatures. Vinit Bansal et. al designed a 3D heater for ZnO gas sensor with a very high temperature (1000 °C) and with very small applied voltage (2 V).

Double meander structure is more suitable and compatible for design of micro-heater on 3D gas sensor structure for uniform temperature distribution with high withstanding capacity [8–11].

2 Experimental

In this work, we have designed a micro-heater for 3D gas sensor with SnO_2 (Stannic Oxide, Rutile Structure) as a sensing material using COMSOL Multiphysics 5.0 simulating tool. Heater based on Joule's heating is constructed on the backside of the glass substrate with nichrome alloy. In comparison with platinum, gold and silver metals for heating materials; nichrome is the more suitable material due to its low manufacturing cost, high thermal conductivity (11.3 $Wm^{-1}K^{-1}$), higher electrical resistivity ($1 - 1.5 \times 10^{-6}$ Ω m), high temperature stability, low-temperature coefficient of resistance (TCR), high resistance to oxidation and no need for extra adhesive layer for power supply contact.

2.1 Design of the Sensor

SnO_2 thin film of 10 μm thickness is layered on 200 μm thick glass substrate in order to support sensor and heater. Heater is built on another side of the glass (SiO_2) substrate at Z = 0.201 cm (i.e., work plane) of the sensor. Geometry of the heater is selected as a meander structure in order to get uniform temperature distribution over all the surface of the sensor. Meander structure is designed with polygon geometry as shown in Fig. 1. Symmetry is maintained in order to get uniform heating over all the substrate. The dimensions of gas sensor are $1.3 \times 0.8 \times 0.201$ cm. The 3D design of SnO2 gas sensor is shown in Fig. 2. The sensor is meshed with free triangular meshing element of a minimum size of 0.02 cm. In order to improve the solver's performance, segregated solvers are created for temperature, displacement and potential distribution.

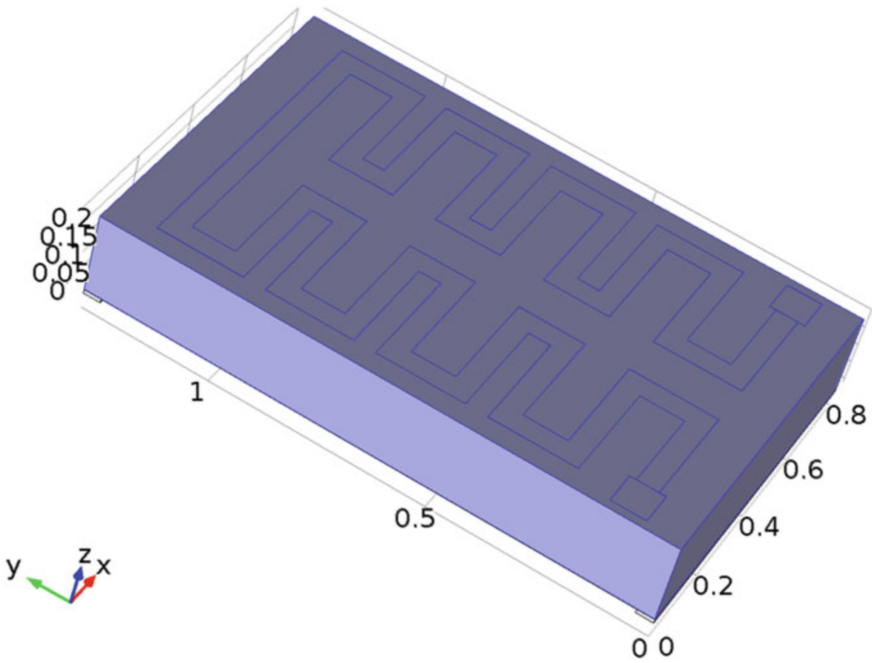


Fig. 1 Geometry of heater

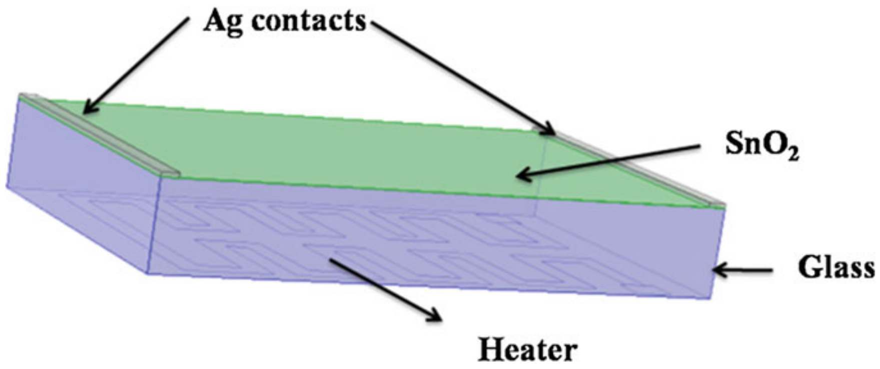


Fig. 2 3D-SnO_2 gas sensor

Table 2 Paramaters defined for simulations [12–14]

Name	Expression	Description
V_in	1.85 (V)	Input voltage
d_layer	10 (um)	Layer_thichkess
Sigma_silver	6.3e7 (S/m)	Electrical conductivity of silver
Sigma_nicrome	9.5e5 (S/m)	Electrical conductivity of nicrome
T_air	20 (°C)	Air temperature
h_air	5 (W/(m^2*K))	Heat transfer film coefficient of air

Nichrome is used as a heating material to design meander shape of 0.05 cm heating coil. Two square-shaped silver contacts of size 0.015 × 0.015 cm were made on the two terminal of the heater. Different voltages were applied to one terminal of the heating coil with respect to another terminal, which is grounded. Table 2 summarizes the parameters fixed for simulation of micro-heater.

In micro-heater simulation, solid mechanics is intended for general structural analysis of 3D bodies and to set initial values, fixed constraints and prescribed displacements. In 3D plane, Navier's equations are used to compute stresses and strains in the gas sensor. The heat transfer in solid's interface is used to model and transfer the heat by conduction, convection and radiation. A heat transfer in solid's model is defined at boundary by setting boundary, initial values and heat flux. Electric field, current and potential distributions are computed by adding electric current's shell interface in thin conducting shells under conditions, where inductive effects are minute. The membrane interface is used to model the pre-stressed membranes and a thin cladding on a solid. Membranes are considered as plane stress in 3D with a possibility to deform both in the in-plane and out-of-plane directions. Finally, all physics are coupled by COMSOL Multiphysics and overall results are computed.

The stationary study is used to compute static electricity, direct currents and temperature field over micro-heater. In solid mechanics, it is used to compute temperature distribution, electric potential distributions, deformations, stresses and strains at static equilibrium.

3 Results and Discussions

At 1.9 V, temperature of the heater was found in the range of 618–621 K (that is 345–348 °C). This temperature range is optimized since SnO_2 gas sensor mostly operates with high sensitivity to gases. The temperature is found to be 348 °C on the coil surface and about 345 °C on the remaining part of the substrate. The temperature profiles with linear distance variation on the substrate are shown in graphs. The inset shows the linear length on different cross sections on the centre of the substrate and another between the central and edge line of the substrate. Potential distribution along the heating coil is shown in Fig. 3. Gradient of the potential is found symmetric, which provides uniform heating in the coil.

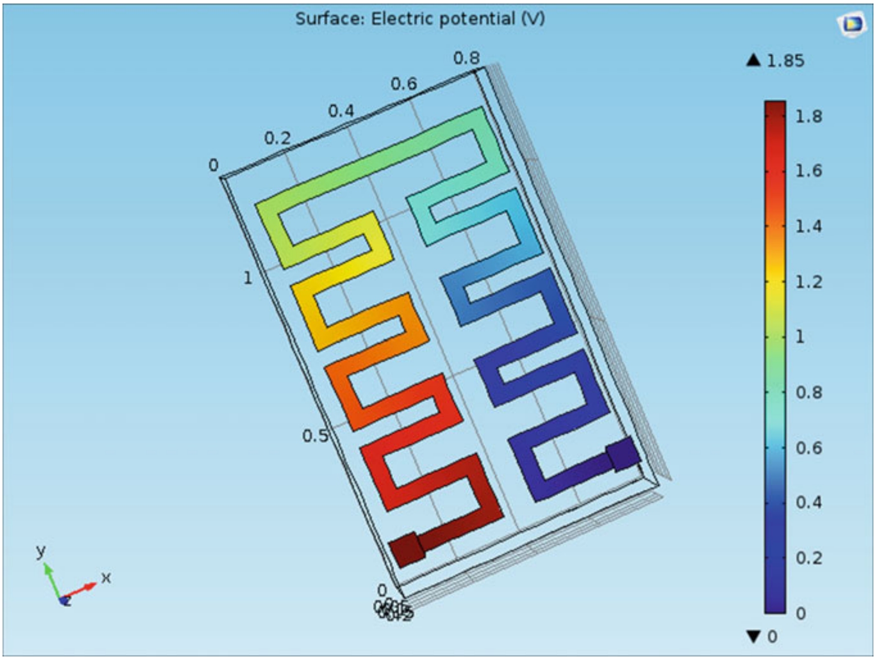


Fig. 3 Potential distribution

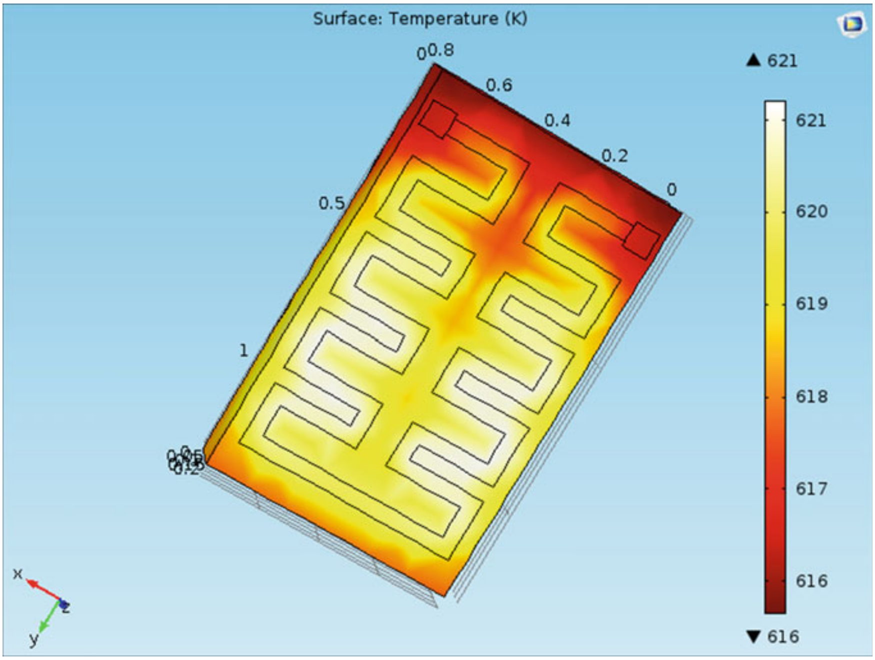


Fig. 4 Surface temperature profile

Heating profile of the gas sensor is shown in Fig. 4. This profile shows quite a good stabilization in the range of 345–348 °C, and hence it is used for sensor. In order to study the variation of temperature on the surface of the sensor, which is shown in Fig. 5, a 3D cut (inset) line is drawn on the surface of the sensor. Overall variation of the temperature on the surface is about 4 °C and the variation near heating element is about ±2 °C, which is observed. This variation in the temperature of the sensor surface is considerable as it does not affect on the sensing property of the sensor since the resistance of SnO_2 is not varying significantly in minor temperature variation.

The von Mises stress on the surface of gas sensor is shown in Fig. 6. The highest stress yield for gas sensor is found nearly 59.7 MPa, which is much smaller than the yield stress of glass (250 MPa) and nichrome (360 MPa). Hence, entire elements of gas sensor stayed intact for heating simulation.

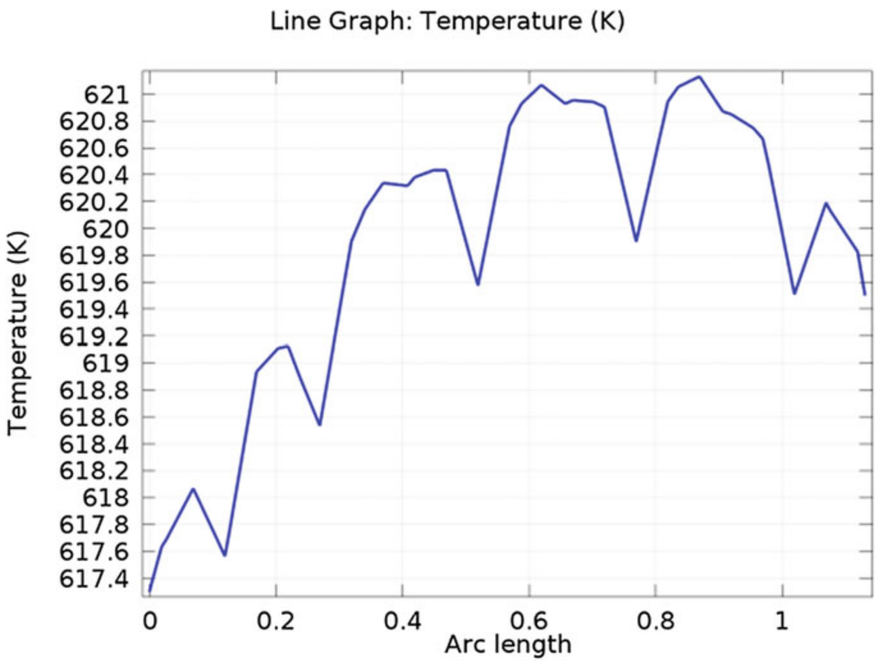


Fig. 5 Temperature Variation with 3D cutline over a heater surface (inset)

In comparison with several reports similar to this work, Monica et al. simulated nickel alloy (Dilver P1) micro-heater with spiral geometry for ZnO gas sensor and reported low-power consumption due to nickel dilver P1 alloy. Marius Dumistrescu et al. simulated poly silicon hot plate on silicon substrate with 3D finite element method and achieved 400 °C temperature by applying 100 mW power. S. Semanic et al. have reported a polysilicon micro hot plate on SiO2 substrate, which is compatible and easily gets integrated on heterogeneous types of chip circuits. J. C. Belmonte reported Ti/Pt micro-heater based BaSnO3 sensor for oxygen and carbon monoxide gases working with 600 and 700 °C [15–18]. The results obtained in this work are compatible with the reported work, and hence nichrome-based micro-heater and 3D-designed sensor model look promising for gas sensing application.

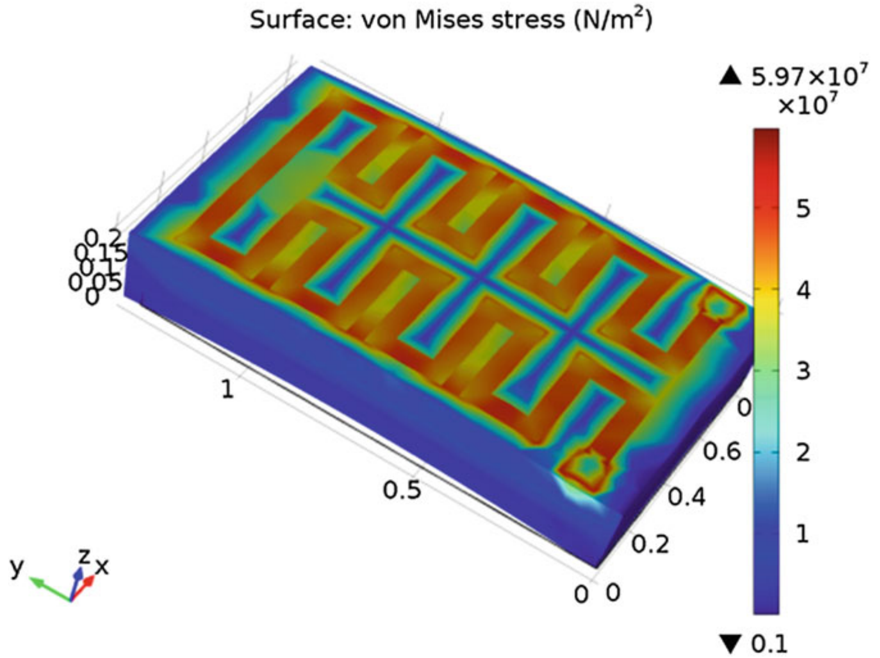


Fig. 6 Surface von Mises stress

4 Conclusion

Micro-heater based on Joule's heating effect is designed on 3D-SnO_2 gas sensor with double meander structure that gives promising results. Nichrome is found to be a more convincing heating element when compared to platinum, gold, and silver in terms of properties and more importantly with cost, which makes the sensor economically cheaper. Uniform temperature distribution with $\pm 2\,°C$ variation on the surface of the sensor is reported. This uniformity of temperature is helpful to increase the stability and accuracy of the sensor. Stress yield of sensor material and glass substrate is quite smaller. This smaller stress yield provides high withstanding capacity of the sensor at a given operating temperature. Simulation of gas sensing response for various gases and selectivity study of the sensor will be studied in the future as it will help to find closer picture of sensing mechanism and also gives a sophisticated protocol to the sensing industry.

References

1. Eason, G., Taguchi, N.: Japanese Patent Application (1962)
2. Batzill, Matthias, Diebold, Ulrike: Prog. Surf. Sci. **79**, 47 (2005)
3. Leo, G., Rella, R., Siciliano, P., Capone, S., Alonso, J.C., Pankov, V., Ortiz, A.: Sens. Actuators B: Chem. **58**(1), 370–374 (1999)
4. Sberveglieri, G.: Recent developments in semiconducting thin-film gas sensors. Sens. Actuators B: Chem. **23**(2–3), 103–109 (1995)
5. Mitra, P., Chatterjee, A.P., Maiti, H.S.: ZnO thin film sensor. Mater. Lett. **35**(1–2), 33–38 (1998)
6. Karunagaran, B., Uthirakumar, P., Chung, S.J., Velumani, S., Suh, E.-K.: TiO_2 thin film gas sensor for monitoring ammonia. In: Materials Characterization, vol. 58 (8–9), pp. 680–684 (2007)
7. Eranna, G., Runthala, D.P., Gupta, R.P.: Oxide materials for development of integrated gas sensors-a comprehensive review. Crit. Rev. Solid State Mater. Sci. (3–4), 111–188 (2004)
8. Umbarkar, S.B., Rajput, G.Y., Vasappanavara, R.: Design and analysis of CO_2 sensor using COMSOL multiphysics. In: Proceedings COMSOL Conference (2016)
9. Velmathi, S., Ramshanker, G., Mohan, N., Design, S.: Electro-thermal simulation and geometrical optimization of double spiral shaped microheater on a suspended membrane for gas sensing. In: Proceedings of the 36th Annual Conference on IEEE Industrial Electronics Society, pp. 1258–1262 (2010)
10. Sujatha, L., Selvakumar, V.S., Aravind, S., Padamapriya, R., Preethi, B.: Design and analysis of micro-heaters using COMSOL multiphysics for MEMS based gas sensor. In: Proceedings of the COMSOL Conference (2012)
11. Bansal, V., Gurjar, A., Kumar, D., Prasad, B.: 3-D design, electro-thermal simulation and geometrical optimization of spiral platinum micro-heaters for low power gas sensing applications using COMSOL. In: Proceedings COMSOL Conference (2011)
12. Dugdale, J.S.: The Electrical Properties of Metals and Alloys. Dover Publications (2016)
13. Swanson, J.G., Campbell, D.S.: The structural and electrical properties of 80: 20 NiCr thin films. Thin Solid Films **1**(3), 183–202 (1967)
14. Matula, R.A.: Electrical resistivity of copper, gold, palladium, and silver. J. Phys. Chem. Ref. Data **8**(4), 1147–1298 (1979)
15. Monika, Dr, Arora, A.: Design and simulation of MEMS based microhotplate as gas sensor. Int. J. Adv. Res. Comput. Eng. Technol. **2**(8), 2487–2492 (2013)
16. Dumitrescu, M., Cobianu, C., Lungu, D., Dascalu, D., Pascu, A., Kolev, S., van den Berg, A.: Thermal simulation of surface micromachined polysilicon hot plates of low power consumption. Sens. Actuators A: Phys. **76**(1–3), 51–56 (1999)
17. Semancik, S., Cavicchi, R.E., Wheeler, M.C., Tiffany, J.E., Poirier, G.E., Walton, R.M., Suehle, J.S., Panchapakesan, B., DeVoe, D.L.: Microhotplate platforms for chemical sensor research. Sens. Actuators B: Chem. **77**(1–2), 579–591 (2001)
18. Cerdà Belmonte, J., Manzano, J., Arbiol, J., Cirera, A., Puigcorbé, J., Vilà, A., Sabaté, N., Gràcia, I., Cané, C., Morante, J.R.: Micromachined twin gas sensor for CO and O2 quantification based on catalytically modified nano-SnO_2. Sens. Actuators B: Chem. **114**(2), 881–892 (2006)

Blackbox-Based Night Vision Camouflage Robot for Defence Applications

Harsh Surana, Nitesh Agarwal, Akash Udaykumar and Rucha Darekar

Abstract Camouflage robot plays a big role in saving human loses as well as the damages that occur during disasters. The main purpose of the paper is to design Blackbox with camouflage robot. One of the main features of this robot is camouflaging, i.e., sensor will catch the image of the surrounding, and the color of the surrounding will be detected by the color sensor and according to that the camouflage robot will change its color. Because of this feature, the robot cannot be easily detected by the enemies. Thus, it will gain more importance in the upcoming era. The robot basically consists of a vehicle mounted with color sensor, which is a part of camouflaging technique and night vision camera is used for observation purpose. Camouflage robot can be sent up to the required area for capturing the unusual happening from attacker. The camouflage robot basically works as an aid for the military. The motion of the camouflage robot can be operated by ZigBee module.

Keywords Camouflage · Surveillance · ZigBee · Blackbox

1 Introduction

A camouflage robot is a technically improved device which can be controlled by any electronic gadget on the basis of human need. A robot is capable of being programmed, it is versatile and worked in such a way to identify any explosive object

H. Surana (✉) · N. Agarwal · A. Udaykumar · R. Darekar
Department of Electronics and Telecommunication Engineering, Pillai HOC College of Engineering and Technology, Rasayani, India
e-mail: harsh.surana1996@gmail.com

N. Agarwal
e-mail: nagarwal@mes.ac.in

A. Udaykumar
e-mail: akashudaykumar1996@gmail.com

R. Darekar
e-mail: ruchadarekar14@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_64

present on the ground, or above the ground. It is a multifunctional device which makes the human work easier and can also perform the task which is probably impossible by a human. This camouflage robot is designed to make the work of military man easier. It can be very handful for protection of our nation against inappropriate/harmful object. The foremost use of this camouflage robot is at the boundary of the nation. We have used many different types of sensing component for the formation of robot. And, the most unique thing of this camouflage robot is that it can be controlled from a longer distance with the help of ZigBee module. Our camouflage robot will be mainly operated by the military, and if unwanted incidents took place then with the help of GSM and GPRS module, the message and exact location of the robot will be send to control room of the military. So that extra help can be given to them [1–6].

Science is developing new technologies to ease human life. One such invention is specialized robots in the field of Artificial Intelligence. The word robot means “A machine that is capable of carrying out a complex series of actions automatically, especially one programmable by a computer”. We have created our device by using two different ideas in which our primary purpose is to show camouflage and the secondary purpose is to show BlackBox [7–10].

1.1 Camouflage

The main motto of the camouflage robot is to minimize human death in defence and spying operation. This camouflage robot can perform essential spying tasks. The function of the camouflage robot is to change its color with the help of color sensing module and according to its surrounding, it gets completely camouflaged, so that it can be impossible to identify it by the bare eyes. The camouflage robot also consists of night vision camera for recording cumulative video/tape. This night vision camera can be very helpful for daytime as well as nighttime.

1.2 Blackbox

In this camouflage robot, the main purpose of BlackBox system is to identify the unappropriate object/matter, harmful gas with the help of gas sensor. The heart rate of the living being can be calculated by counting the number of beats which can be recorded/identified by heart rate sensor. This can give us a result about the current situation of military man whether he is dead or alive. We have also used night vision camera for observation purpose.

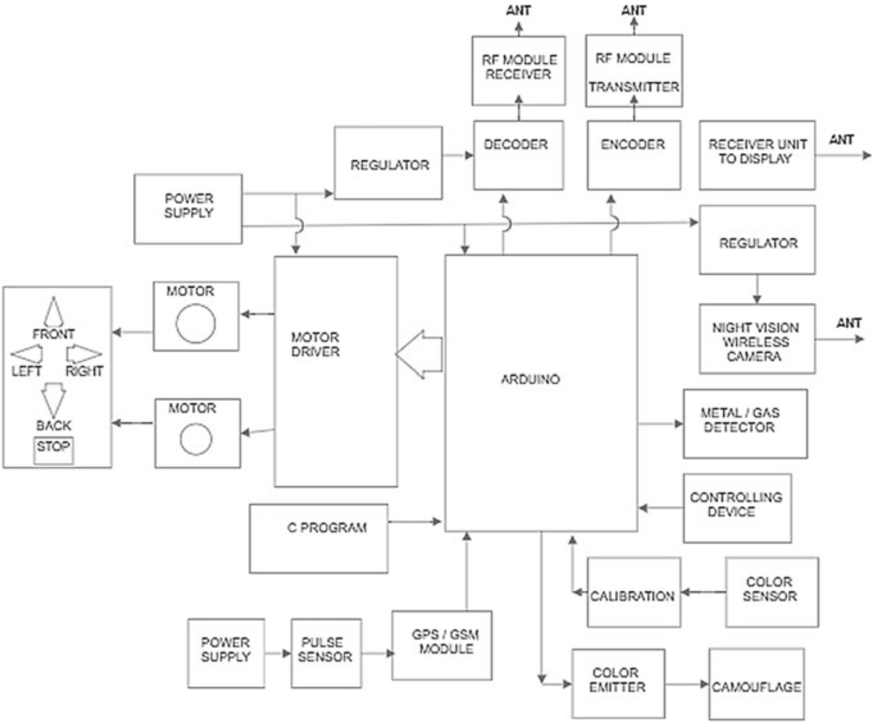


Fig. 1 Block diagram of camouflage robot

2 Block Diagram

See Fig. 1.

3 Hardware Specifications

3.1 Zigbee CC2500

In this paper, control of robot is from remote end, and also we are able to get the videos from the robot end for the purpose of surveillance. Night vision camera can help us to get the live video of the situation, so that the military can take immediate action according to the situation. DC motors are being used for the movement of robotic wheels and for camera movement, i.e., upward and downward movement. For wirelessly transmitting commands to the robot ZigBee transceiver, i.e., CC2500 is used.

3.2 Color Sensor-TCS34725

For the camouflaging technique to take place, the essential component which we are going to use is color sensor-TCS34725. This component consists of RGB and clear light sensing element. To reduce the IR spectral component of the incoming light and to allow color measurement to get the accurate image, an IR blocking filter, and integrated on-chip is used. This will produce much more clear/filtered image of the situation. Since humans cannot see IR. The other benefit of using this color sensor is it occupies less space, and can be easily implemented.

3.3 Heart Rate Sensor

The heart rate of the living being can be calculated by counting the number of beats which can be recorded/identified by heart rate sensor. If it is not possible to calculate the heartbeat, we can use the technique of calculating the pulse rate by simply touching fingertip. In this camouflage project, we are using the pulse rate technique which is much more efficient and time-saving, and can be easily performed with the help of a robot.

3.4 Arduino Uno R3

Microcontroller can be used to develop interactional object, taking inputs from various sensors, and controlling a variety of light, motor, and other outputs. There are many different microcontrollers available which include Arduino, Raspberry Pi, depending on the brand and version, a microcontroller can consist of different peripheral and battery voltage. Depending on the power requirement and the type of the sensor, a microcontroller is selected. In the present work we have used Arduino Uno as the controller.

3.5 GSM Module-SIM900A

The major role of GSM (Global System for Mobile) in this camouflage robot is to send the message on the electronic device of current happening to the military control room. It is suitable for SMS, Voice as well as data transfer. GSM, modem-rs232 is built with a dual-band GSM engine. SIM900A works mainly on frequency 900/1800 MHz.

3.6 GPS Module-SIM28M

The main function of Global Positioning System is to track the location of the place where the incident is taking place. So that military control room can get the exact position of current happenings.

3.7 Night Vision Camera

We have used a 1.2G wireless night vision camera Radio AV Receiver which can be very helpful for daytime as well as nighttime. This camera is small in size and is basically used for observation purpose.

4 Results and Discussions

The aim of the paper is to design a camouflage robot with advanced design of BlackBox system which operates via any electronic gadget like smartphone, and laptop is used as a remote to control the device. We have used a ZigBee CC2500 module, so that we can control the robot from a longer distance; the maximum area which will be covered is about 1.8 km approximately. The primary purpose is to show a camouflage technique and the secondary purpose is to show BlackBox. For the camouflage part, we have used a color sensor-TCS34725, this sensor will catch the image of the surrounding, and according to that the camouflage robot will change its color. To achieve this goal, we have used a common anode LED, which will display RGB color depending on the surrounding. A mini night vision camera is also used which will give the live video with the help of Radio AV receiver. On the other hand, we have created a BlackBox which consist of various sensors like gas sensor to examine the harmful toxic gases, GSM module is used to get the message of current happening, GPS module to track the exact location, heart rate sensor to calculate the pulse rate of military man which is done with the help of pulse rate technique, and LCD display to display the detected parameter. This camouflage robot will work as an useful aid for the militaries as shown in (Figs. 2 and 3).

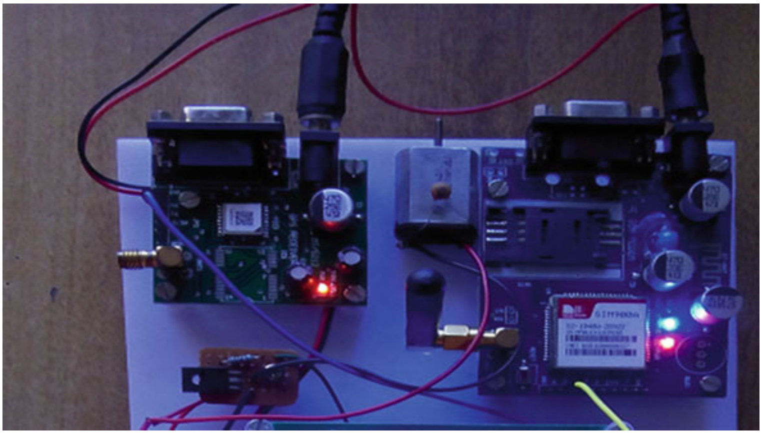


Fig. 2 Connecting GSM and GPS module

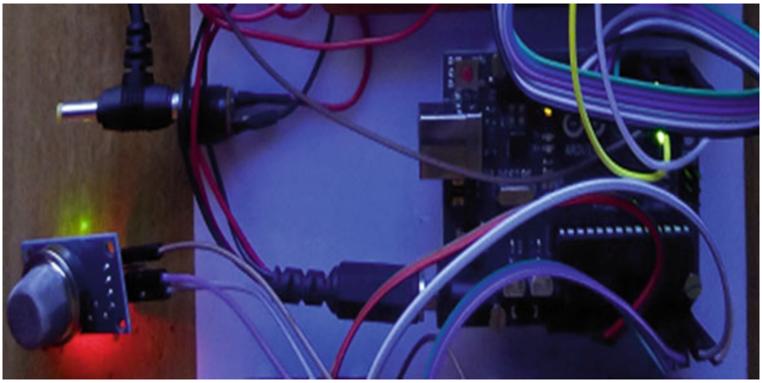


Fig. 3 Connecting the sensors

5 Conclusion

The proposed system provides a helping hand to our security forces in the detection of intruders. The robot can also be used in high altitude areas where a human cannot survive. Moreover, the camouflaging feature makes it difficult to detect the robot by naked human eye. There is scope to improve the system by configuring it with multicolor camouflaging.

References

1. Schultz, J., Drury, J.L., Yanco, H.A.: Evaluation of human-robot interaction awareness in search and rescue. IEEE 2327–2332 (2004)
2. Premkumar, M.: Unmanned multi-functional robot using ZIGBEE adopter network for defense application. Int. J. Adv. Res. Comput. Eng. Technol. **2**(1) (2013)
3. Hymavathi, P., Jyothy, T.: Surveillance alive human detection robot using Zigbee technology. Int. J. Electron. Commun. Eng. (SSRG-IJECE) (2014)
4. Ravindran, A., Premkumar, A.: Camoflage technology. Int. J. Emerg. Technol. Comput. Sci. (IJETCSE) **8**(1) (2014). ISSN:0976-1353
5. Bekey, George: Autonomous Robots: From Biological Inspiration to Implementation and Control. MIT Press, Cambridge, MA (2005)
6. Arun Kumar, M., Sharmila, M.: Wireless Multi Axis ROBOT for Multi-Purpose Operations, Department of ECE, SVCET. JNT University Anantapur, India (2006)
7. Rashid, R.A., Zhen, H.L.: Wireless monitoring system using Zigbee IEEE 802.15.4 Standard. In: International Wireless Telecommunication Symposium (IWTS06) (2006)
8. Bhargavi, S., Manjunath, S.: Design of an Intelligent Combat Robot for War Fields, Department of Electronics and Communication Engineering, SJCIT, Chikballapur, Karnataka, India (2006)
9. Kapoor, C., Kannan, G.K.: Landmine detection technologies to trace explosive vapour detection technique. Def. Sci. J. **57**(6), 797–810 (2007)
10. Naghsh, A.M., Gancet, J., Tanoto, A., Roast, C.: Analysis and design of human-robot swarm interaction in firefighting. In: Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, Technische Universität München, Munich, Germany (2008)

IOT-Based Wi-Fi Surveillance Robot with Real-Time Audio and Video Streaming

Diksha Singh and Anil Nandgaonkar

Abstract In this paper, the wireless robot refers to the mini robot which live streams the monochromatic video, takes and stores the images. The robot is being controlled through a local Wi-Fi server by a compatible web page. The objective of the proposed method is to implement the aforementioned technology pertaining to the mini robot, which is capable of performing multiple tasks at an affordable cost. Arduino Uno R3-Based Robot Control Board will be used to design the robot and. In this paper, we have proposed a surveillance robot with the facility of real-time video streaming, audio transfer, and ability to avoid obstacles in the process. The system will be designed as such to stream the video live to the person monitoring the robot. We have used two Android phones in the proposed method for the purpose of video streaming and audio transfer. An entire new approach for controlling the robot through web page has been used. We have used NodeMCU ESP Module, to incorporate wireless connectivity in the proposed method.

Keywords Wi-Fi robot · NodeMCU · IPCAM app · L293D motor driver

1 Introduction

The term robot found its origin in a Czech word having the meaning slave. A robot basically is an intelligent device designed to help humans in almost every relevant or irrelevant field. The robot is capable of performing even a Herculean task. The robots have no fixed shape, neither they have been specified for any particular field or for any particular work. They can be made or converted into any form depending upon the area of application. With the advancement of wireless communication technology

D. Singh (✉) · A. Nandgaonkar
Department of Electronics and Telecommunication, Dr. Babasaheb Ambedkar Technolgical University, Lonere, India
e-mail: singhdeeksha1993@gmail.com

A. Nandgaonkar
e-mail: abnandgaonkar@yahoo.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_65

in mobile robots, there is a great possibility, wherein we shall have a cell phone controlling a robot maid, wheelchair, or an autonomous robot car. Surveillance is the process of monitoring the particular location. This mostly happens in the areas where continuous monitoring of the adversary territory is essential for the nation's security. It is not possible for human beings to go to any area they want. Human access is restricted in certain blocked off place. Moreover, loss of life is also feared in the case of getting caught in the adverse territory. However, with the recent advancement in technology, this dangerous task of surveillance is performed by the robots instead of human beings. With zero chance of human loss and with immense ease they provide, several surveillance robots have been developed all around the world up till now.

In the proposed system, a new surveillance system for indoors is added as additional features to transmit voice from mobile surveillance is developed where it can be transmitted wirelessly to robot and surveillance area to mobile phone. The robot can be moved in all the four directions (left, right, front, and back) using its surveillance area effectively. The mobile phone is used to give the commands to the robot to move in all the four directions and receive the video. The communication between the wireless surveillance robots and the mobile phone is done by the internet. Above is the general block diagram of the surveillance robot. We are using 12v DC MOTOR for driving the robot. The revolution per second (rpm) of DC Motor is 100. In DC motor, we have more rpm but not exact angular movement like servo motor. For controlling the motor, we are using L293D Servo Stepper DC Motor Shield to control the movement of 12 V DC motor (Fig. 1).

2 Literature Review

Some of the recent works done in field of surveillance robotics is mentioned. Wren et al. [1] have analyzed about real-time tracking of the human body. Shah et al. [2] have discussed about an automated surveillance system, which is used in a range of real-world places starting from security in the railway to the enforcement of the law. Ramya et al. [3] have presented an embedded system for surveillance robot using ZigBee and web server as well. The robot uses gas sensors for leakage of the gas and for the detection of the intruders. The robot is controlled from a remote area by the use of Zigbee technology. Borker et al. [4] have designed a surveillance robot which uses an Android smartphone to control the robot. The robot consists of the microcontroller, motor drivers, and motors. The robot is capable of surveillancing along with transmission of the video. Kadiam et al. [5] have discussed about a Smartphone-Controlled Two Axes Robot for Video which uses Wireless Internet and Raspberry Pi Processor. The robot can be made to be moved in all the four directions using an Android application. The video surveillance is done with the help of a robotic vehicle which is interfaced to the USB along with arm 11 processor. Pramod et al. [6] have presented a wireless robot systems using Arduino microcontroller. The communication between the mobile and robot is through the ZigBee protocol which restricts the coverage range of the robot. Irsahina et al. [7] have discussed about an

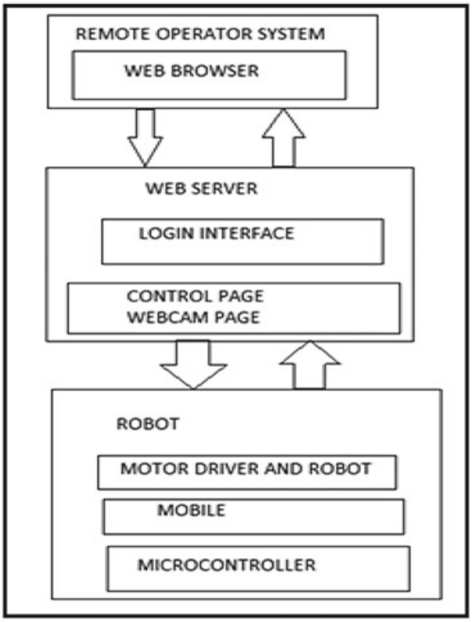


Fig. 1 General block diagram of surveillance robot

Arduino-based robot which is designed using Bluetooth technology to control robot and Wi-Fi network to transmit the video. Letian et al. [8] have discussed about a robot which performs the image processing through the camera mounted on the top. Sri et al. [9] have presented a surveillance robot for home security.

3 System Objective

The objective is to use robot for surveillance system over Wi-Fi network and to transmit the audio and images as well in the process. The whole Transfer Process will be performed Real Time. The surveillance robot will be controlled by the Android smartphone using the internet. The robot can move in all the four directions using the robot control web page available on the phone or the computer. The key motivation is to provide a more cheap and effective surveillance system, which can be used in the dangerous areas or the place where it is almost impossible for the human beings to go. Since the NodeMCU Wi-Fi module and Android Mobile needs to be paired using username and password, which is only known and predefined by the programmer in the programming, hence this Wi-Fi robot control is much secured from intruders. It has many functional features like

3.1 This system gives four functional commands (Forward, Reverse, Turn left, and Turn right) to control the Wi-Fi bot.
3.2 It will be able to record and transmit the live video. It will have the ability to transmit the live audio and it can transfer the images as well
3.3 We will be using a software app called as ***IPCAMERA*** for the purpose of the live transmission purpose. Upon entering the IP address in any favorite browser, the page of the IPCAMERA opens up.
3.4 A secure Login System can also be created in the proposed system so that only authorized person can access and control the movements of the robot. We have not created database instead included the username and password in the code itself.
3.5 We will be designing an HTML web page for controlling the movements of the robot. The generated page will have five buttons namely forward, reverse, left, right and stop.

In this proposed method, the role of IOT comes into the picture when I am able to control my robot from anywhere. I am using HTML server page for controlling the robot. Through port forwarding, I can control my robot from anywhere in the world. However, in my proposed system, I am assigning a specific IP address so that both the NodeMCU and the remote for controlling the robot should be on the same network.

4 Hardware Overview

The proposed surveillance robot consists of the following components.

- L293D Motor Driver.
- NodeMCU ESP8266 Wi-Fi Module.
- DC Motors.
- Android Mobile containing IP Webcam.

4.1 L293D Motor Driver

L293D is a typical Motor driver or Motor Driver IC which allows DC motor to drive on either direction. L293D is a 16-pin IC which can control a set of two DC motors simultaneously in any direction. It means that you can control two DC motors with a single L293D IC. The l293d can drive small and quiet big motors as well (Fig. 2).

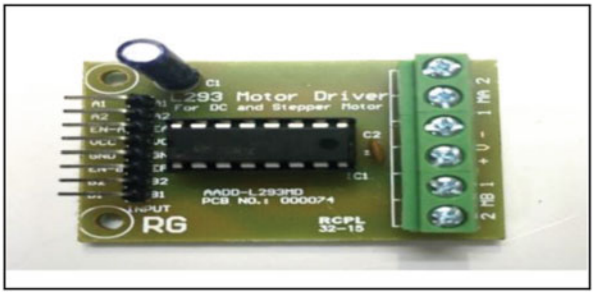


Fig. 2 L293D motor driver

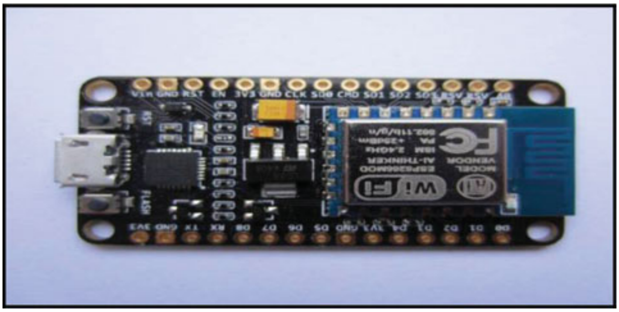


Fig. 3 NodeMCU ESP8266 Wi-Fi module

4.2 NodeMCU

NodeMCU is an open source IoT platform. It includes firmware which runs on the ESP8266 Wi-Fi SoC from Espressif Systems, and hardware which is based on the ESP-12 module. The NodeMCU ESP Wi-Fi 8266 module provides a minimum of 512 Kb flash memory. It is low-cost, user-friendly plug-and-play module with easy to configure and set up. It is widely used to develop hardware platform in IOT application. This device is also called as mini Arduino. It has its unique IP address (Fig. 3).

4.3 Android Phone Having IP Webcam Software

IP webcam turns your phone into a network camera with multiple viewing options. It has the capacity to stream video inside Wi-Fi network without internet access. It is a great app that uses your old phones camera to capture and send streaming video directly to any web browser in the world. All you have to do is to download, install it, and run the app with the phone with the camera facing the area you wish

Fig. 4 IP webcam

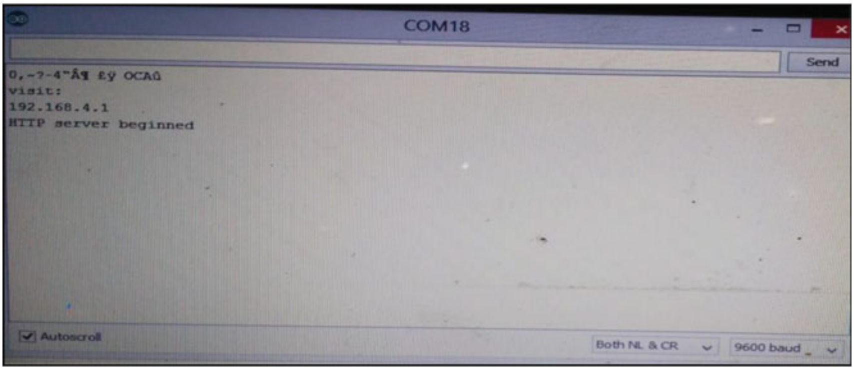


Fig. 5 The generated IP address

to monitor. You will need to enter your Wi-Fi's connection password and jot down the IP address the app gives you for monitoring the video/audio and images. Then, visit that IP address with the browser on any connected computer or smartphone to view the action (Fig. 4).

5 Software Part

As far as the proposed system is concerned, the web page is generated by connecting to the hotspot of the NodeMCU. There is no requirement of the internet for generating the web page and that is the biggest advantage here. Upon uploading the program in NodeMCU, a specific IP address will be generated (Fig. 5).

Upon entering the IP address in the browser, the below-given web page opens up.

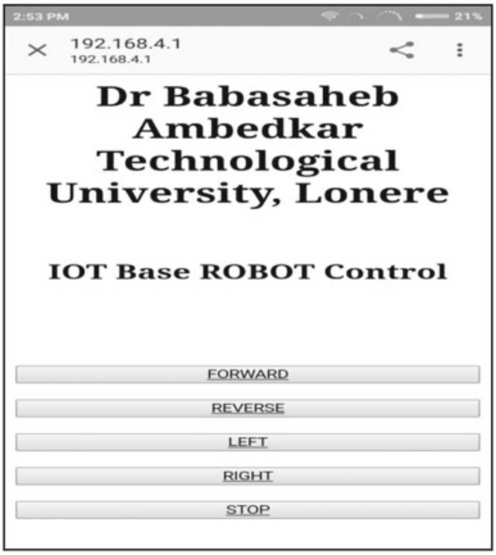


Fig. 6 Web page for controlling the robot

6 Results and Discussion

The resulting system after applying the proposed method will be able to live stream the video, audio as well as image transfer facility will also be there. Since the usage of hardware and wiring is minimum, possibility of breaking down of the robot is reduced to minimum. As far as working principle of the robot is concerned, we will be sending the respective commands, i.e., forward, backward stop, and start via the web page of the host device be it PC or the Android mobile (web page is already shown in Fig. 6). We will be keeping the baud rate of NodeMCU as 115200 bps. Although in the proposed system, we have used NodeMCU only for controlling the robot, but for further modifications, we can use Arduino Uno R3-based control board also. Since NodeMCU has only one analog pin. The speed at which the motor will move can be specified in the program and can be varied between 0 and 255. We have proposed to control the robots direction by creating a GUI.

6.1 Video Streaming Section

We are proposing to use an old Android phone as camera in this system. For real-time video, audio streaming section, I am using IP webcam. I have IP webcam installed in the phone which I will be using as a camera. As soon as we start the IP webcam, the IP address along with the port will not be generated. We can input the generated

IP address in our favorite browser and see the real-time video streaming, audio streaming, and image transfer page opening up.

7 Concluding Remarks

In the proposed method, we have tried to have a new approach for making the surveillance robot. Robots movement is being controlled by the IP address and generated web page for controlling the robot movements. Various technologies have emerged so far in the field of surveillance. Infrared and Radio frequency has been used for the surveillance and detection of the human body and intrusion detection. In the budding years of robotics technologies such as Bluetooth and ZigBee had been used to control the robot from a distant or remote place. We observed that in earlier days that the camera mounted on top of the robot for video processing and image processing was fixed and could not be moved in any other direction. In certain robots, the night vision camera has also been used for capturing of the image at night. PIR and IR sensors are also deployed in the robots for moving object detection and for intrusion detection. The robots can avoid the obstacles as well and can transmit the videos and audio live. With the advancement of technology solution such as using the SD card and rechargeable battery, has evolved leading to the increase in the efficiency of the system. Surveillance robots using Wi-Fi technology instead of the regular Bluetooth and ZigBee technology have arrived, capable of covering large distance at a lower cost. Hand gesture-based surveillance robots have also been developed which are particularly useful for physically handicapped and elder people. This proposed system has an entirely new approach and instead of the traditional Arduino Uno, NodeMCU has been used. It acts as a mini Arduino and a Wi-Fi module as well. We will be using L293D motor driver for the purpose of controlling the motors. NodeMCU documentation will be used for the programming purpose. We will be using web page for controlling the movements of the robot. The Real-Time Video, Audio, and Image Transmission have been done through the IPCAM app installed in my Android phone, which we will be used as a camera. We will be using android phones both for the real-time transmission of the data and for controlling the movements of the robot.

References

1. Wren, C., Ajarbayejani, A., Darrel, T., Pantland, A.: P finder, real time tracking of human body. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 780–785 (1997)
2. Shuguang, C.U.I.: Secured Spy IP Control Robot Using Raspberry Pi. IEEE J. Sel. Areas Commun. **22**(6), 1089–1098 (2004)
3. Shah, M., Javed, O., Shafique, K.: Automated visual surveillance in realistic scenarios. IEEE **89**(10), 1355 (2007)

4. Ramya, V., Palaniappan, B.: Web based embedded robot for safety and security applications using ZigBee. Int. J. Wirel. Mob. Netw. **4**(6), 155–174 (2012)
5. Pramod, S.V., Srikanth, N., Vivek, M.U., Sarat, P.C.: Intelligent intrusion detection system (In2DS) using wireless sensor network. IEEE **8**(7), 56 (2012)
6. Liu, L., Zhu, X., Tang, Y.: Indoor surveillance robot controlled by a smart phone. In: Proceeding of the IEEE, China, pp. 1875–1880 (2013)
7. Piyare, R.: Internet of things: ubiquitous home control and monitoring**** systems using android based smart phone. Int. J. Internet Things **2**(1), 5–11 (2013)
8. Kadiam, V., Pavani, G.: Smart phone controlled two axes robot for video, using wireless internet and raspberry Pi processor. Int. J. Res. Advent Technol. (IJRAT) **2**(10), 97–100 (2014)
9. Borker, K., Gaikwad, R., Rajput, A.S.: Wireless controlled surveillance robot. Int. J. Adv. Res. Comput. Sci. Manag. (IJARCSMS) **2**(2), 436–441 (2014)

Genetic Algorithm Approach for Obstacle Avoidance and Path Optimization of Mobile Robot

Sunil B. Mane and Sharan Vhanale

Abstract The path planning is an important issue of mobile robots. Its task is to find a collision free path from the start position to the target position with an algorithm which requires less time and minimum path distance. The scheduling and planning is NP-Hard (NP-Complete) problem. Autonomous robot vehicles can be used in variety of applications including space exploration, household and transportation. In known static environment path planning algorithms such as Sub Goal network, A* algorithm, D* Star algorithm, Artificial Potential Method are used. These are classical and heuristic search based algorithms. The above mentioned algorithms have some drawbacks such as local minima, deadlock of robot, and oscillation of robot. We have proposed an algorithm which will overcome these drawbacks present in existing classical algorithms.

Keywords Artificial intelligence (AI) · Artificial neural network (ANN) Genetic algorithm (GA)

1 Introduction

Mechatronics has main role in the field of robotics. But while navigating and path planning of mobile robot performance is not satisfactory. This is due to inability of robot to gather real time information and its processing. To overcome this machine

S. B. Mane (✉)
Department of Computer Engineering and Information Technology,
College of Engineering, Pune, MH, India
e-mail: sunilbmane.comp@coep.ac.in

S. Vhanale
Department of Production Engineering and Industrial Management, College of Engineering,
Pune, MH, India
e-mail: sharnvhanale@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_66

vision with artificial intelligence can be used in robots. This will help robots in decision making for obstacle detection and avoidance. Additionally genetic algorithm will improve the decision making based on optimum path which is calculated based on its cost.

2 Literature Survey

Mobile robot navigation and obstacle avoidance using many AI techniques involves the various approaches such as static obstacles and moving obstacles. Till now ANN is applied to static environment with static obstacles in mobile robot navigation.

Howard Li et al. [1] developed an intelligent planner which automatically generates possible set of actions out of which the desired one is selected for robots in a workspace with moving obstacles. This system is similar to biological inspired neural network, which builds a neural network containing several neurons which are characterized by a shunting neural equation. This type of neural network based path planning has a low level of navigation control on robot which is not behavioural. This proposed system tested in simulation environment only. Guanghui Li et al. [2] represents a method of regression search which uses an improved APF.In this method the problem of local minima and oscillations is addresses by obtaining a global suboptimal path efficiently in complete known environment. The simulations results show the efficiency of improved APF method resolving the problems of local minima and oscillatory movements. Song-Hiang Chia et al. [3] proposes the Ant Colony System Based Mobile Robot Path Planning which uses the ant colony algorithm for optimizing the path of mobile. To find the path the four vertices of obstacles are considered which are used to program the motion path using ant colony algorithm. This system does not guarantee the path computed is optimized or not. Also this algorithm is tested for static environment only. Youssef Bassil [4] gives an ANN network in which a multilayer structure is developed which consists of three layers: an input layer, a hidden layer, and an output layer. This proposed ANN network is trained by use of back-propagation supervised learning algorithm in offline mode for static environment only. Simulation is used to carry out the neuron training. The execution time of training process is higher as is does not uses the parallel and distributed approach for training the neurons. Frantisek Duchon et al. [5] summarize the A* algorithm and its modifications by keeping in mind optimal path in less execution time.In some experiments this algorithm gives shortest time for finding the path but the obtained path is longer as compared to standard eight connectivity algorithms. Hence, this algorithm is feasible to use in cases where it is necessary to quickly find a path. Roy Glasius et al. [6] developed a model of an organized neural network. The analytical results obtained show the performance of neural network. This system depends on large collection of neurons, which increases the computation time. The quality of path is somewhat degraded and network becomes instable as number of neurons increases.

3 Background

3.1 Back Propagation Algorithm

Many researchers suggests artificial neural network for mobile robot navigation and path planning. Back-propagation algorithm is commonly used in neural network training. The main steps involved in back-propagation training are as follows:

1. Give input and desired output to the neural network.
2. Compute output based on activation function with error.
3. Check for error. If error is within acceptable range then take next input or go back to the pervious stage to update network weights.
4. Repeats this step until all inputs are given to the neural network.

Additionally for controlling the performance of the network learning rate and momentum are used. The basic idea behind for setting these parameters is trial and error method. For some values of momentum and learning rate, neural network gives an optimum solution of weights and bias values, but for some values it cannot give the optimum solution. By increasing the learning rate we can speed up the process of getting converging solution but it gives to unstable network. So to achieve desired momentum and weight values we need a trial and error method which can provide the stable network. It is the main weakness of back-propagation. For this purpose we are using genetic algorithm training.

3.2 Genetic Algorithm

Genetic algorithm is a meta-heuristic technique for solving searching and optimization problems. In this a set of new population is generated from the fitness value of previous generation. It is based on phenomenon of natural selection, where an initial set of population is generated randomly and fitness value of each solution is computed by applying the fitness function of the problem. Then the fit solutions are selected and genetic operations such as mutation and crossover are used to get new population of fit individuals. These steps are repeated until the maximum numbers of generations are reached or a desired fitness value is attained. GA is used successfully by researchers for solving searching and optimization issues like travelling salesman problem, job scheduling.

3.3 Problem Statement

Lot of work has been done in area of path finding and optimization of path in robot navigation area of robotics for static environment. But the existing works have not

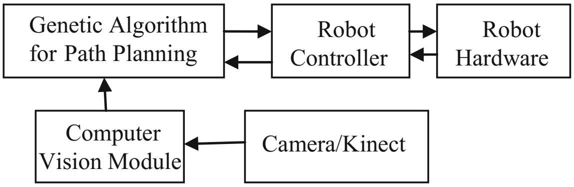


Fig. 1 System architecture

addressed the issue of dynamic nature of the environment satisfactorily. Hence there is a scope in this area to explore various possible solutions to such dynamically changing environment such as moving objects, uncertain obstacles in the path, etc. There is a need of having an intelligent path finding algorithm which can take care of such dynamically changing environment. Hence there is need for design an algorithm for computer vision to extract desired information such as environment and obstacles and a genetic algorithm for finding shortest path and reduce the required traversal time.

3.4 Proposed System Design

The proposed system design is divided into four modules as shown in Fig. 1.

3.4.1 Computer Vision Module

Image is acquired using Kinect. The depth map images obtain using OpenCV. Depth map is converted into top view. Then this view is divided into two parts upper part and lower part depending on the height of sensor. The vertical line in image is represented as obstacles. Hough transform is used to detect vertical lines. Threshold of top view image is computed. Edge detection algorithm is applied to get the boundary lines of obstacles. Kalman filter are used to stabilize the effect of fluctuations.

3.4.2 Path Planning Module

The map is created with image captured with camera vision module and given as inputs to the genetic algorithm. We are proposing our method in MATLAB simulation. For simulations purpose we creating map image that contains static obstacles of different sizes. Image map is considered as search space and it is divided into 100 × 100 grids. Initial position is considered as (0, 0). Goal position is at (100, 100).

3.4.3 Robot Controller

The robot controller module takes decision to guide the robot in environment. Depending on the obstacle position robot controller gives commands to motor driver to move in the environment. Robot controller takes actions such as move forward, move right, move left to guide the robot for optimizing the path.

3.4.4 Robot Hardware

The robot hardware module moves in real world environment following the commands received by robot controller. Robot reaches to the specified target location, avoiding the obstacles and following the optimum path computed by GA.

4 Implementation of System

Implementation of proposed system is divided into four modules. We tasted our simulation results with MATLAB and the system having configuration Intel i3, 2.4 GHz 2.4 GHz with 6 GB RAM. Simulation results show the optimal and collision free path for mobile robot. Modelling of the system includes finding a path from start position to target position. The obstacles are considered as point in real world. The path is computed by joining these points by a straight line. So the objective function becomes the path length of robot.

4.1 Initialization of Population

The initial population size is collection of all points in grid that are perpendicular to the x-axis. Initial population is generated randomly. The computation time and accuracy of the path depends on the population size. Larger the population size larger the computation time but path accuracy is more. Our algorithm mainly identifies the via-points in the grid that represent a path from start location to goal position.

4.2 Fitness Function

For modelling the system we need an objective function which is to be optimized. While designing GA fitness function plays an important role to stabilize the algorithm and to get optimum solution. Fitness function is nothing but the summation of each evaluation function. In our problem statement, we have two problems one to determine the obstacles in grid and second one is to optimize the path. We can use a single function but it leads to instability of genetic algorithm.

Fitness function for path optimization is normally Euclidean distance formula which given by,

$$fit2 = \sum_{i=0}^{n-1} \sqrt[2]{\left((x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2\right)} \tag{1}$$

The path obtained from this function is a straight line. To get smooth path we are using spline function. The fitness value is computed for each generation.

4.3 Genetic Operators

4.3.1 Initial Population

Global optimal solution and computation time is depends on population size. The population size indicates the no of paths in the grid. In general the size of population is defined in range of [20, 100]. In our problem statement population size is defined as 50.

4.3.2 Selection

In this step the individuals are selected for reproduction depending on fitness values. Other are discarded. There are many methods for selection process such as fitness proportionate selection and tournament selection, etc. We are using tournament selection. The best individuals are selected for crossover operation.

4.3.3 Crossover

In this step the selected parents are recombined to produce child depending on the uniform crossover operation.

4.3.4 Mutation

To avoid the problem of local minima we are introducing the mutation operation. In this operation the small probability is added to improve the previous solutions. Generally the mutation probability ranges between [0.001, 0.4]. In our algorithm we are defining our mutation rate as 0.2.

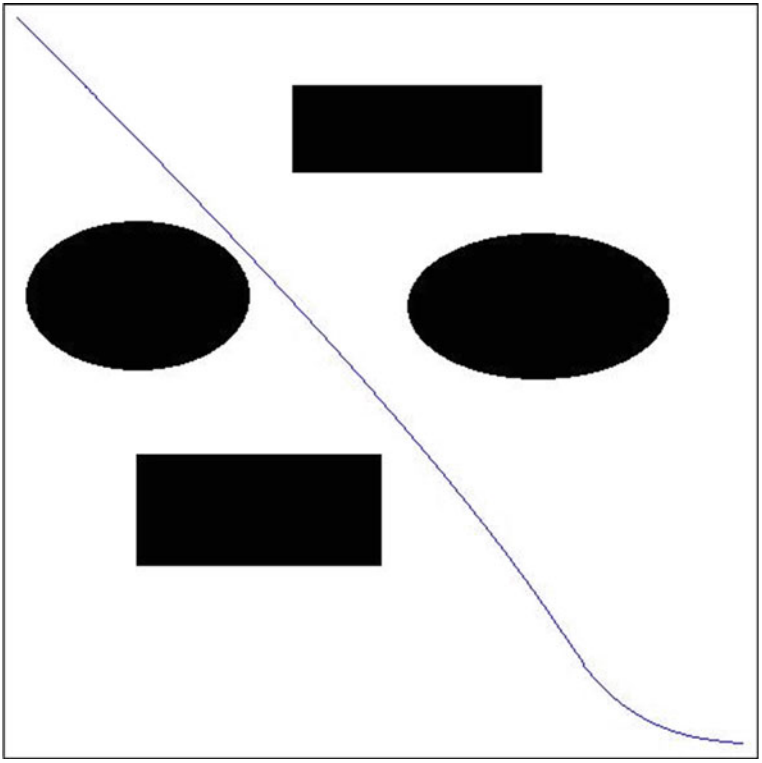


Fig. 2 Optimize collision free path For Map1

5 Experimental Results

For experiments we have created three maps containing various numbers of obstacles of different sizes and shapes. We are analyzing the performance of genetic algorithm by increasing the number of obstacles. The main objective of the optimization problem is to minimize the fitness value after each generation. More diverse population means more optimum solution. Figure 2 shows collision free path in environment containing four obstacles of different sizes and shapes. The start position is at (0, 0) and target position is at (100, 100) which is not moving. The circles and rectangles are considered as obstacles in the environment. The fitness value of the objective function is gradually decreasing after each generation as shown in Fig. 3. After 50 generations the fitness value is tending to the zero value. The convergence of genetic algorithm is verified by observing the average distance between individuals in population as shown in Fig. 4. The results are also tested on the various maps by increasing the number of number of obstacles.

Figure 5 shows a map containing five obstacles. The best fitness values for this map are taking values near to the zero but the path cost is increased as the numbers of

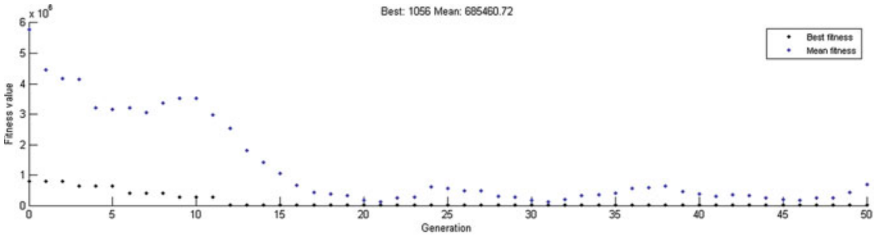


Fig. 3 Best fitness values for Map1

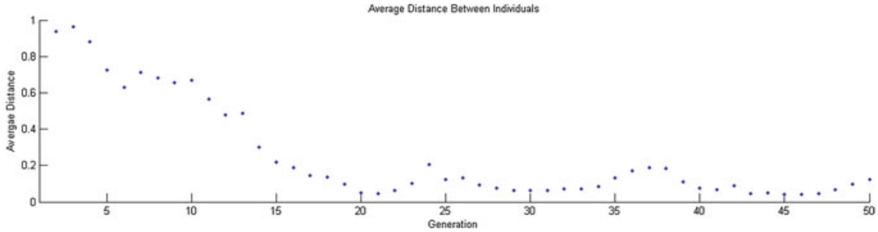


Fig. 4 Average distance between individuals for Map1

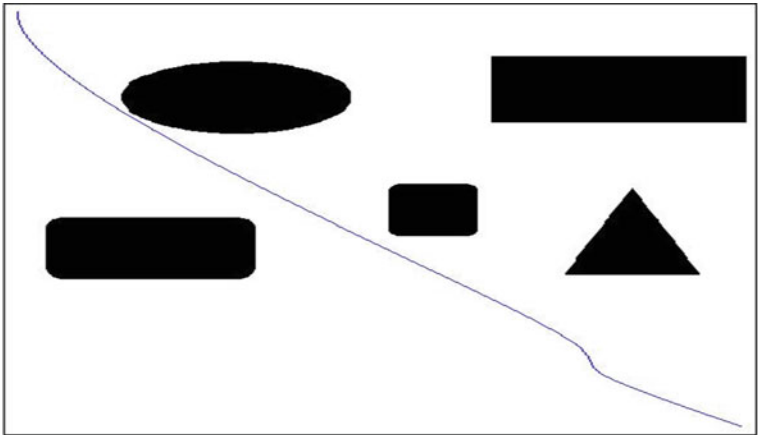


Fig. 5 Optimize collision free path for Map2

obstacles are increased as shown in Fig. 6. Figure 7 shows the more optimal solution as compare to the graph of Map1.

Similarly for map3 the path cost is increased because of the increasing number of obstacles. As the numbers of obstacles increased the genetic algorithm takes some unnecessary turns as shown in Fig. 8, which will add to the total path cost mobile robot. The graph of best fitness values and the average distance between individuals shows the optimal solution for map3 (Figs. 9 and 10).

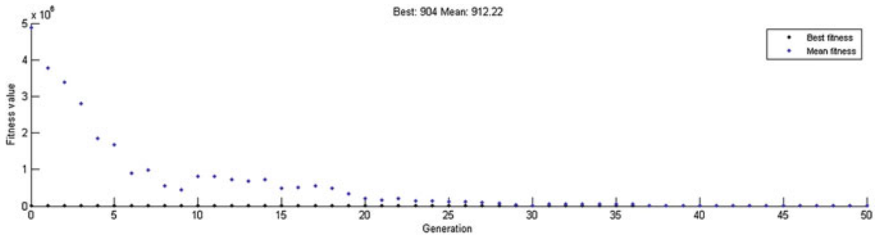


Fig. 6 Best fitness values for Map2

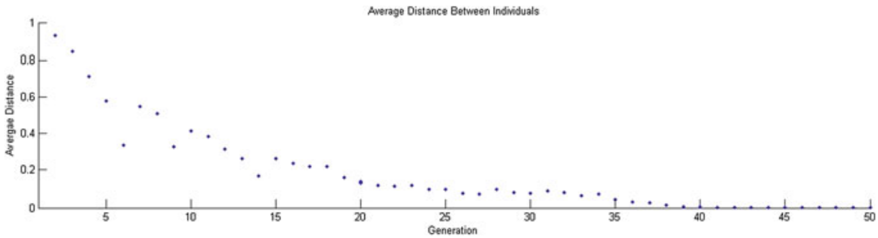


Fig. 7 Average distance between individuals for Map2

Table 1 Simulation results

Sr. no	Simulation results		
	No. of obstacles	Processing time(s)	Path length
Map1	4	93	1276
Map2	5	99	929
Map3	7	97.1	1203

Total path traversal cost and time required to compute path between starting position to goal position is given in Table 1.

6 Conclusions and Future Work

The proposed algorithm gives optimal collision free path with respect to distance, changing environment, and traversal time for mobile robot. Proposed algorithm is designed using genetic algorithm and real-time image processing techniques. The distance between individuals of population in one generation is directly proportional to the convergence of the genetic algorithm. The path cost and traversal time increases as the number of obstacles in the environment increases. We have used fifty generations to execute our proposed algorithm. The fitness value of objective function is gradually decreases in each generation of proposed genetic algorithm which will give

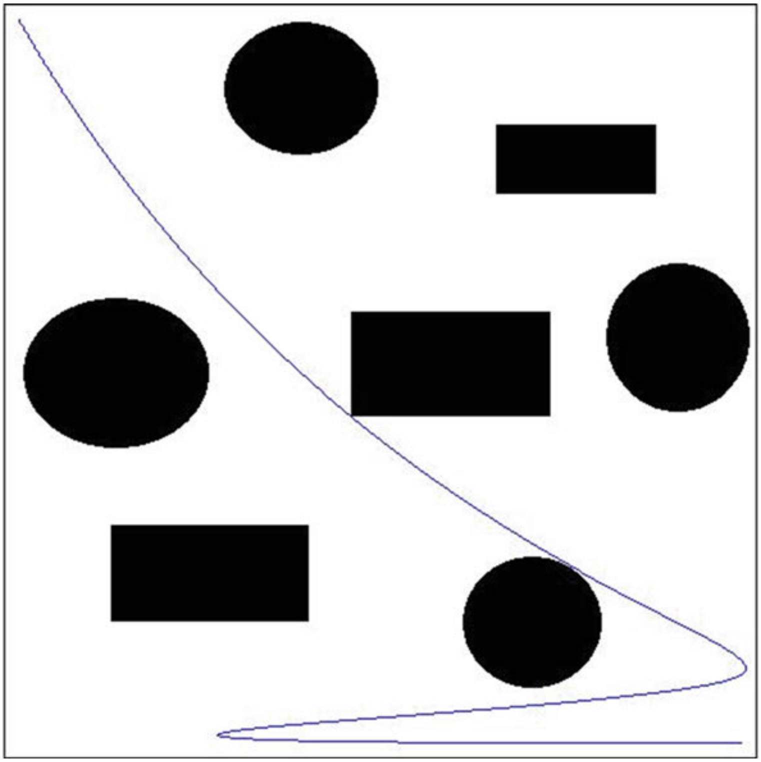


Fig. 8 Optimize collision free path for Map3

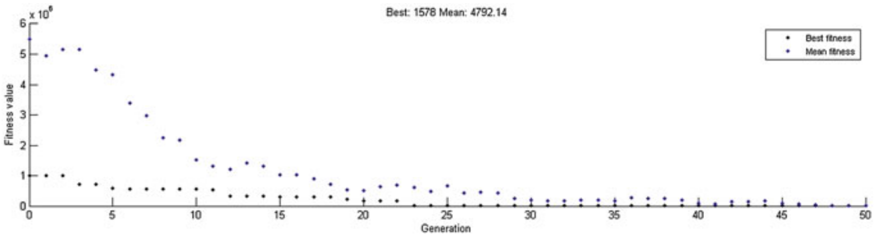


Fig. 9 Best fitness values for Map3

optimum solution in the final generation. This method can also be used for dynamic environment with moving target by making relevant changes in proposed algorithm.

Future work involves the use of GA for ANN training which evolves the network of ANN and gives weight and bias values.

This algorithm can be used real world environment using hardware realisation of the robot like pi-bot.

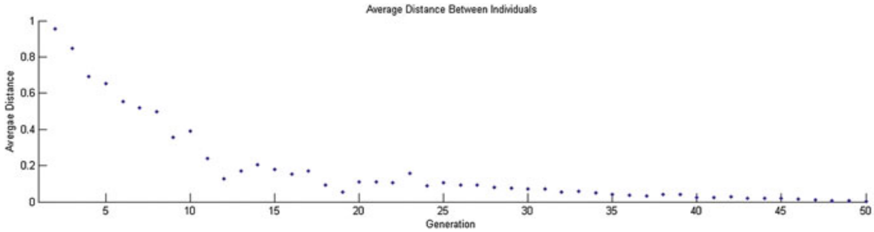


Fig. 10 Average distance between individuals For Map3

References

1. Li, H., Yang, S.X., Seto, M.L.: Neural-Network based path planning for a multi-robot system with moving Obstacles. In: IEEE transactions on Systems and Cybernetics Part C: Application and Reviews, vol. 39, no. 4, July 2009
2. Li, G., Yamashita, A., Asama, H., Tamura, Y.: An Efficient Artificial Potential Field Based Regression Search Method for Robot Path Planning
3. Chia, S.-H., Su Jr., K.-L., Guo, H., Chung, C.-Y.: Ant colony system based mobile robot path planning. In: 2010 Fourth International Conference on Genetic and Evolutionary Computing (2010)
4. Bassil, Y.: Neural network model for path _planning of robotic rover. Int. J. Sci. Technol. (2012)
5. Duchon, F., Babinec, A., Kajan, M., Beno, P., Florek, M., Fico, T., Jurisica, L.: Path planning with modified A* Algorithm for a mobile robot. Model. Mech. Mechatron. Syst. MMaMS (2014)
6. Glasius, R., Komoda, A., Gielen, S.C.A.M.: Neural network dynamics for path planning and obstacle avoidance. Neural Netw. **8**(1),125–133,195
7. Guldner, J.: Autonomous Navigation System for a Mobile Robot or Manipulator United States Patent, Patent Number: 5,758,298

Performance Verification of DC–DC Boost Converter

Vaibhav Marne and K. Vadirajacharya

Abstract The DC–DC boost converters had marked their importance in the field of renewable energy sources with their excellent features. Along with the time, various topologies of boost converters have been introduced to enhance its efficiency and make it more reliable. These boost converters are associated in many applications along with certain drawbacks and a research trend is developed to improve the performance of boost converters. This paper carries out the simulation of the proposed converter and tries to prove the effectiveness of the converter that overcomes the drawbacks of many boost converter topologies such as low voltage gain, high-input ripple current, high duty ratio, high inductor core losses and high stress on switch. The simulation has been performed on PSIM tool and the results of the proposed converter are verified using PSIM software.

Keywords Quadratic boost converter · Ripple current · Flying capacitor
Continuous conduction mode

1 Introduction

In the recent years, the advancement in power electronics had introduced DC–DC boost converters in the field of power systems. The application of these converters in renewable energy sources has greater benefits in both technical and economical ways. However, employing these converters creates certain limitations. To overcome these limitations and enhance the performance of DC–DC boost converters, various researches and studies are carried out in the field of power electronics.

V. Marne (✉) · K. Vadirajacharya
Department of Electrical Engineering, Dr. Babasaheb Ambedkar Technological University, Lonere, India
e-mail: marnevaibhav4@gmail.com

K. Vadirajacharya
e-mail: kvadirajacharya@dbatu.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_67

The conventional boost converter had made tapping of LC branch for ripple current cancellation. The inductor tapping reduces the input ripple current. However, the gain of this converter is the same as CBC [1]. The lower dc voltage from PV system is boosted efficiently using three-level boost converter. Here, as the level of the converter is increased the capacitor voltage must be in balanced condition, which is achieved by using phase-shift method. The perturbation and observation method is used to achieve maximum power [2]. However, the voltage gain is lower. In [3], a zero-ripple input current boost converter is used for high-gain applications and a clamp capacitor is connected in series to transfer energy to the load, so the converter can produce a very high voltage and high conversion efficiency. Non-isolated high-gain converter has better performance as compared to CBC and three-level boost converter in terms of gain, voltage stress and switching losses. This converter is having inherent voltage balancing [4]. The capacitor and inductor were applied in switched manner in Mc-SC/SL-SBC that increases the output voltage gain. Here, the size of converter is reduced but as the number of cells is increased the duty ratio is reduced and the converter becomes more complex [5]. Quadratic boost converter's output voltage is almost double the input voltage with advantages like reduced hardware and low switching losses [6]. In [7], flying-capacitor boost converter is having high voltage gain ratio and improved conversion efficiency that can be achieved with low-inductor core and copper losses. The FCBC has high gain as compared to QBC.

2 Operating Principle and Steady-State Analysis of Proposed Converter

Figure 1 shows the circuit diagram of quadratic flying-capacitor boost converter. The analysis of converter is done under steady state. The converter is composed of two switches S1 and S2 with a switching period of T. At input side of the converter, the two inductors L1 and L2 that oppose the input ripple current are present. The capacitors C1, C2 and C3 charge and discharge according to switching of the switches. Diodes D1, D2, D3 and D4 get forward biased when the anode is more positive than cathode.
Switching state 1: The switches S1 and S2 are closed for period t, $\left[0 < t < \frac{DT}{2}\right]$. Inductors L1 and L2 get charged during this period. The diode D2 is forward biased due to the inductor voltage VL1 and diodes D1, D2 and D3 are reverse biased as shown in Fig. 2.

Switching state 2: The switch S1 is open for period t, $\left[\frac{DT}{2} < t < \frac{T}{2}\right]$. Inductors L1 and L2 pass their stored energy to capacitors C1 and C2, respectively. Diodes D1 and D3 conduct due to forward-biased condition and diodes D2 and D4 become reverse biased as shown in Fig. 3.

Switching state 3: The switches S1 and S2 are closed for period t, $\left[\frac{T}{2} < t < \frac{(1+D)T}{2}\right]$. Inductors L1 and L2 pass their stored energy to capacitors C1 and C2, respectively. Diodes D1 and D3 conduct due to forward-biased condition and diodes D2 and D4 become reverse biased as shown in Fig. 3.

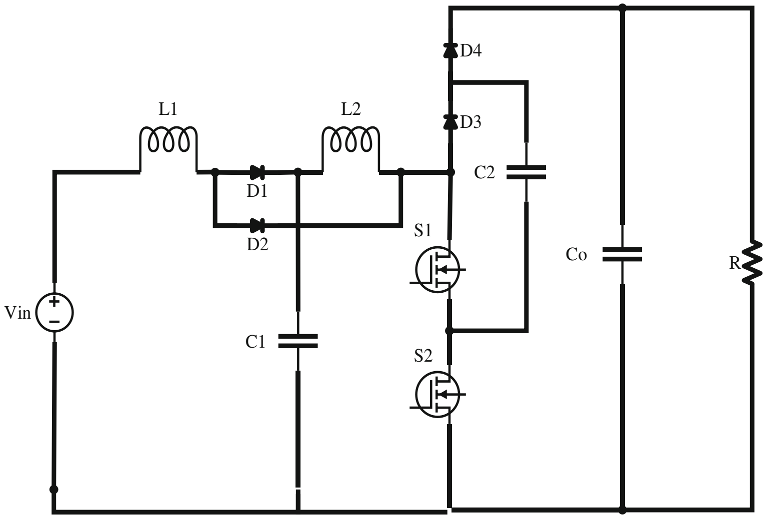


Fig. 1 Quadratic flying-capacitor boost converter

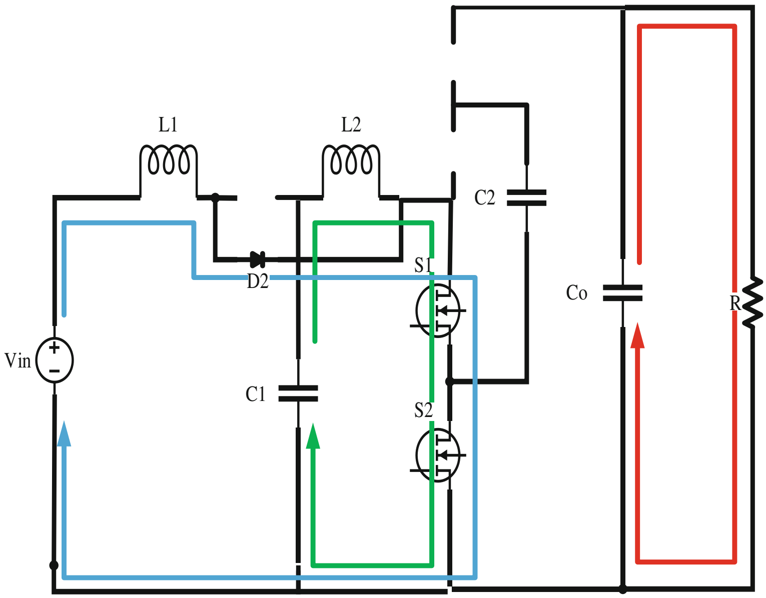


Fig. 2 QFCBC with both switches closed

Switching state 4: The switch S2 is open for period t, $\left[\frac{(1+D)T}{2} < t < T\right]$. The stored energy in inductor L1 is delivered to capacitor C1 and stored energy of L2 is passed to C0 and load along with the capacitor C2 as shown in Fig. 4.

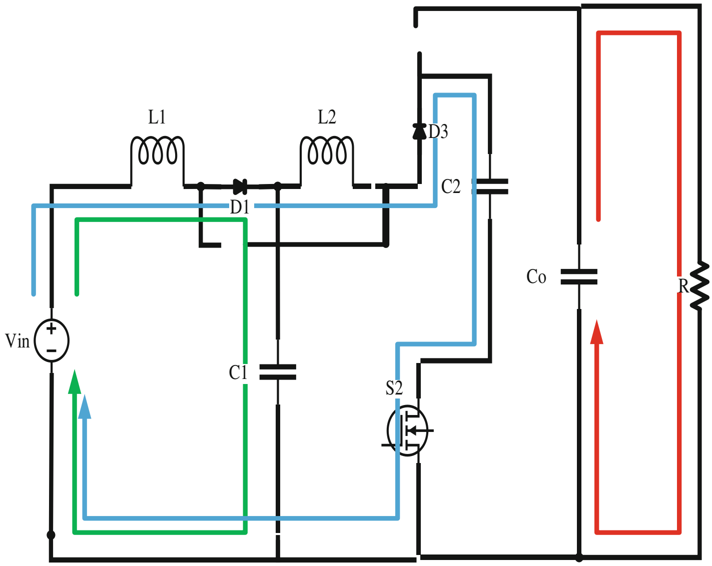


Fig. 3 QFCBC with S1 switch open

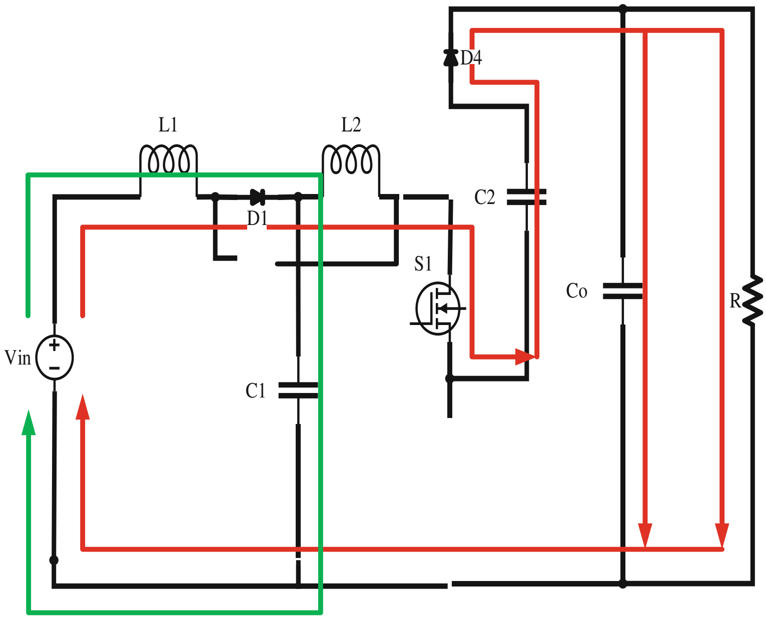


Fig. 4 QFCBC with switch S2 open

3 Performance Verification

3.1 Design Example

The performance verification is done using PSIM software, and Fig. 5 shows the simulation performance in PSIM. For analysing the performance of QFCBC, the parameters considered are $Vin = 48$ V; $Vo = 384$ V; $fs = 10$ kHz; $Io = 6.4$ A; and the design parameters of passive components are known by carrying out the mathematical modelling on the proposed converter. The design values of passive components are listed in Table 1.

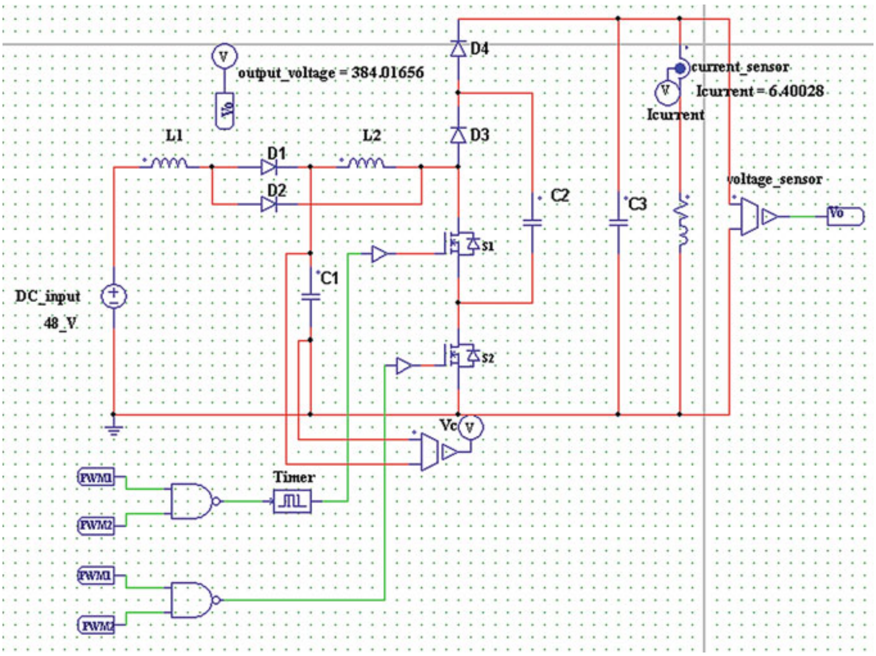
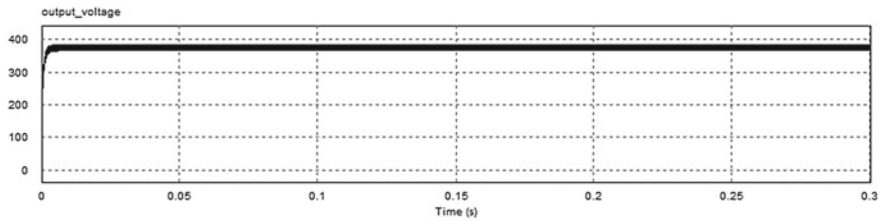


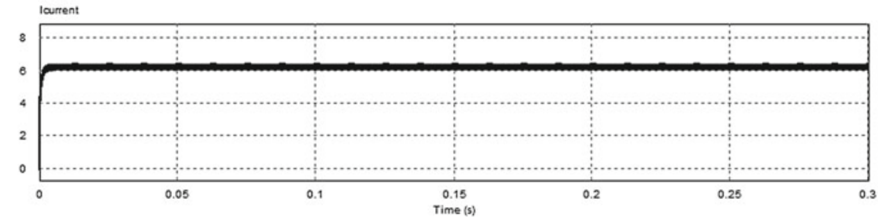
Fig. 5 Simulation of QFCBC in PSIM software

Table 1 Values of passive components

Inductor	Value
L1	11.7 uH
L2	46.87 uH
Capacitor	Value
C1	4 uF
C2	4 uF
C3	25 uF



(a). Output voltage waveform.



(b). Output current waveform.

Fig. 6 **a** Output voltage waveform. **b** Output current waveform

3.2 Simulation Results

Figure 6a, b shows the simulation results of the QFCBC. Figure 6a shows the output voltage waveform, the output voltage graph shows the high voltage gain at low duty ratio and the output voltage has low oscillations due to specific value selection of passive components. Figure 6b shows the output current waveform having an average value of 6.14A and due to the voltage stiff load, the waveforms of voltage and current are similar but have different magnitudes.

4 Conclusion

In this paper, the simulation of quadratic flying-capacitor boost converter has been carried out using PSIM software and the results suggest the effectiveness of the proposed converter in terms of voltage gain at low duty ratio and low-ripple input current. As compared to the conventional boost converters such as the quadratic boost converter and flying-capacitor boost converter, the proposed converter has low switching stress and low inductor core losses.

References

1. Chang, C.-C., Lee, C.-C., Huang, J.-C., Chen, Y.-M.: Ripple current cancellation for boost converters. 978–1-5090-5157-1/17/$31.00 ©2017 IEEE
2. Tampubolon, M., Lin, W.-C., Lin, J.-Y., Hsieh, Y.-C., Chiu, H.-J.: A Study and Implementation of Three-level Boost Converter with MPPT for PV Application. 978-1-5090-5157-1/17/$31.00 ©2017 IEEE
3. Hu, X., Gao, B., Wang, Q., Li, L., Chen, H.: A Zero Ripple Input Current Boost Converter for High Gain Applications, 21 June 2017. IEEE
4. Kang, H., Cha, H.: Member, IEEE, A new non-isolated high voltage gain boost converter with inherent output voltage balancing. IEEE Trans. Ind. Electron. https://doi.org/10.1109/tie.2017.2736508
5. Zhu, B.Z., Senior Member, IEEE, Li, Z., Li, H., Member, IEEE, Ran, L., Senior Member, IEEE, Extended switched-boost DC-DC converters adopting switched-capacitor/switched inducto cells for high step-up conversion. IEEE J. Emerg. Sel. Top. Power Electron. https://doi.org/10.1109/jestpe.2016.2641928
6. Samuel, R.B.M.E.R.,*a, Deepa M.E b, S.Jothivel M.Ea.: A closed loop control of quadratic boost converter using PID-controller. IJE Trans. B Appl. **27**(11), 1653–1662 (2014)
7. Hamidreza, K., Toliyat, H.A.: Student Member, IEEE Fellow, IEEE. Flying-Capacitor Boost Converter. 978-1-4577-1216-6/12/$26.00 ©2012 IEEE

Comparison of Multiple Attribute Decision-Making Methods—TOPSIS and PROMETHEE for Distribution Systems

S. G. Kamble, K. Vadirajacharya and U. V. Patil

Abstract Distribution system (DS) is considered the weakest link in the power system with 5–13% technical losses. Utilities are under pressure to trim down these losses, improve reliability, and power quality of supply to consumers in the deregulated, competitive environment. This has attracted researchers again for reconfiguration with many alternatives available for decision-making such as losses, power factor, voltage profile, cost, and reliability indices like SAIFI, SAIDI, CAIFI, AENS, etc. Multi-Attribute Decision-Making (MADM) is one such popular method available for decision-making which deals with problems through a number of qualitative and quantitative criteria in reconfiguration. In this paper, MADM methods like TOPSIS and PROMETHEE are proposed for finding the compromised best configuration by considering loss minimization, and reliability indices from available alternatives. Two examples are furnished in this paper to show the effectiveness of the methods.

Keywords Distribution system reconfiguration · Loss minimization
Multi-attribute decision-making · TOPSIS · PROMETHEE

1 Introduction

Distribution system (DS) delivers electrical power to the end users and is the first interface of the utility with the consumers. Due to deregulation and competition in distribution sector, utilities are under pressure to minimize operational cost by reducing losses and improve reliability to enhance the overall performance.

S. G. Kamble (✉) · K. Vadirajacharya
Dr. Babasaheb Ambedkar Technological University, Lonere, India
e-mail: sachinkamble80@rediffmail.com

K. Vadirajacharya
e-mail: kvadirajacharya@dbatu.ac.in

U. V. Patil
Government College of Engineering, Karad, India
e-mail: patil_uv@yahoo.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_68

Distribution feeders contain a number of switches that are normally closed (sectionalizing switches) and switches that are normally open (tie switches). The configuration of a distribution network can be modified by changing the status of the tie and sectionalizing switches. The process of changing the topology by altering the open/closed status of the switches is called reconfiguration and can be used to improve the operating conditions of the system [1–5, 13–18]. The distribution system is reconfigured for the purpose of loss minimization, load balancing on the feeders, relieving overloads, maintenance and also affects voltage profile, reliability, power factor, etc. After reconfiguration of distribution system, number of alternatives is available (number of switching combinations) for the decision makers (DMs). The attributes available for DM are losses, voltage profile, power factor, reliability indices (such as SAIFI, SAIDI, CAIFI, CAIDI, and AENS) [6], cost, etc. In this paper, MADM [7–10] methods like SAW, WPM, TOPSIS, and PROMETHEE are used for finding the optimal configuration by considering loss minimization, reliability indices, etc., from available alternatives for practical distribution systems [11, 12].

2 Multiple Attribute Decision-Making (MADM) Methods

In general form, decision table in MADM methods have alternatives, attributes with weight of each attribute. The most commonly used multi-criteria decision-making techniques are Weighted Sum Method (WSM), Weighted Product Method (WPM), AHP (Analytic Hierarchy Process), TOPSIS (For the Technique for Order Preference by Similarity to Ideal Solution), ELECTRE (For Elimination and Choice Translating Reality), PROMETHEE Preference Ranking Organization Method for Enrichment Evaluations), VIKOR (VIšekriterijumsko KOmpromisno Rangiranje), etc.

2.1 Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) Method

The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) method was developed by Hwang and Yoon. TOPSIS method is based on the idea that the selected alternative should have the minimum Euclidean distance from the best solution and the maximum from the worst solution.

The procedure of decision-making for solving distribution system problem using TOPSIS method [7] is as follows:

Step1: Identify and short list the alternatives on the basis of the identified criteria.
Step2: Prepare a decision table or decision matrix.

Step 3: Obtain normalized decision matrix, Rij, by using the following expression.

$$\mathrm{Rij} = \mathrm{mij}/\sqrt{\sum\nolimits_{j=1}^{M} \mathrm{m}_{\mathrm{ij}}^{2}} \tag{1}$$

Step 4: Decide weights of the different attributes as per decision makers preference.
Step 5: Obtain the weighted normalized matrix Vij by using following equation.

$$\mathrm{Vij} = \mathrm{wj}\,\mathrm{Rij} \tag{2}$$

Step 6: Obtain the ideal (best) and negative ideal (worst) solutions. The ideal (best) and negative ideal (worst) solutions can be expressed as

$$\mathrm{V+} = \left\{ \left(\underset{\mathrm{i}}{\overset{\max}{\frac{\mathrm{Vij}}{\mathrm{j}}}} \right), \left(\underset{\mathrm{i}}{\overset{\min}{\frac{\mathrm{Vij}}{\mathrm{j}}}} \right) /\mathrm{for\ i} = 1,\ 2,\ \ldots,\ \mathrm{N} \right\} \tag{3}$$

$$\mathrm{V+} = \{\mathrm{V1+, V2+, V3+}, \ldots.., \mathrm{VM+}\} \tag{4}$$

$$\mathrm{V-} = \left\{ \underset{\mathrm{i}}{\overset{\min}{\frac{\mathrm{Vij}}{\mathrm{j}}}}, \left(\underset{\mathrm{i}}{\overset{\max}{\frac{\mathrm{Vij}}{\mathrm{j}}}}, \in \mathrm{J}' \right) /\mathrm{fori} = 1, 2, \ldots, \mathrm{N} \right\} \tag{5}$$

$$\mathrm{V-} = \{\mathrm{V1-, V2-, V3-}, \ldots.., \mathrm{VM-}\} \tag{6}$$

where J = (j = 1, 2, …, M)/j is associated with beneficial attributes and
J′ = (j = 1, 2, …, M)/j is associated with non-beneficial attributes.

- Vj+ indicates the ideal (best) value of the attribute.
- Vj− indicates the negative ideal (worst) value of the attribute.

Step 7: Obtain the separation measures by the following equations.

$$Si+ = \sqrt{\sum\nolimits_{j=1}^{M} (Vij - Vj+)^2}, \mathrm{i} = 1, 2, \ldots., \mathrm{N} \tag{7}$$

$$Si- = \sqrt{\sum\nolimits_{j=1}^{M} (Vij - Vj-)^2}, \mathrm{i} = 1, 2, \ldots, \mathrm{N} \tag{8}$$

Step 8: Performance index, Pi, can be calculated as follows.

$$\mathrm{Pi} = \mathrm{S} - /(\mathrm{S} - +\mathrm{S}+) \tag{9}$$

Step 9: A set of alternatives is made in the descending order according to the value of performance index Pi.

2.2 Preference Ranking Organization Method for Enrichment Evaluations (PROMETHEE)

The Preference Ranking Organization Method for Enrichment Evaluations (PROMETHEE) was developed by Brans and falls in the category of outranking methods. In this method, alternatives are compared pairwise for each criterion to decide the strength of preference of one over the others.

The procedure of decision-making for solving distribution system problem using PROMETHEE method [7] is as follows:

Step 1: Identify and short list the alternatives on the basis of the identified criteria.
Step 2: Prepare a decision table or decision matrix.
Step 3: Get the information on the decision maker preference function. Let Pi, a1a2 be the preference function associated to the attribute bj.

$$\mathrm{Pi, a1a2} = \mathrm{Gi[ci(a1) - ci(a2)]} \tag{10}$$

$$0 \leq \mathrm{Pi, a1a2} \leq 1 \tag{11}$$

Preference index Ga1a2 can be calculated as the weighted average of the preference functions Pi.

$$\prod \mathrm{a1a2} = \sum_{\mathrm{i=1}}^{\mathrm{M}} wi \ \mathrm{Pi, a1a2} \tag{12}$$

Step 4: Calculate the leaving flow $\varphi^{+}(\mathrm{a})$.

$$\varphi^{+}(\mathrm{a}) = \sum_{\mathrm{x}\,\varepsilon\,\mathrm{A}} \prod \mathrm{xa} \tag{13}$$

Step 5: Calculate the entering flow $\varphi^{-}(\mathrm{a})$.

$$\varphi^{-}(\mathrm{a}) = \sum_{\mathrm{x}\,\varepsilon\,\mathrm{A}} \prod \mathrm{ax} \tag{14}$$

Step 6: Calculate the net flow $\varphi(\mathrm{a})$.

$$\varphi(\mathrm{a}) = \varphi^{+}(\mathrm{a}) - \varphi^{-}(\mathrm{a})$$

Step 7: Decide the ranking based on the scores of net flow.

Table 1 Example from Ref. [11]

Solution	Losses	SAIFI	AENS
1	139.5513	1.1048	0.4422
2	139.9780	1.0327	0.4118
3	141.9160	1.0173	0.4056
4	142.4292	1.0162	0.4054
5	146.2891	1.0042	0.3998
6	146.5133	1.0031	0.3995
7	146.6658	1.0021	0.3999
8	148.6078	0.9982	0.3991
9	150.2031	1.0003	0.3984
10	150.2483	0.9991	0.3982
11	150.9774	0.9910	0.3952
12	152.5900	0.9871	0.3943
13	156.0999	0.9847	0.3936
14	161.5802	0.9841	0.3935

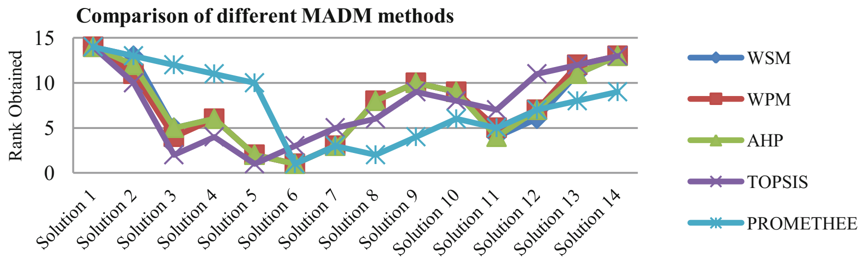


Fig. 1 Graphical representation of comparison of various MADM methods for example [11]

3 Results and Discussions

Distribution System Case Study 1

The example of 12.66 kV, 33-node system is taken into consideration as a case study [11]. This has many possible radial configurations, but only non-dominated solutions are taken as shown in Table 1. The weights of the attributes considered in are 0.3 for the active power losses, 0.35 for system average interruption frequency index (SAIFI) (failures/year), and 0.35 for average energy not supplied (AENS) (kWh/customer/year). The available alternatives for DM are 14, and attributes considered are 3, active power losses, SAIFI and AENS, all the attributes are required to be minimized for the benefit of the distribution system (Tables 2) (Fig. 1).

Result of TOPSIS Method

(See Tables 2,3, 4, 5, 6 and 7).

Table 2 Normalized values

Solution	Losses	SAIFI	AENS
1	1.0000	0.8907	0.8899
2	0.9970	0.9529	0.9556
3	0.9833	0.9674	0.9702
4	0.9798	0.9684	0.9706
5	0.9539	0.9800	0.9842
6	0.9525	0.9811	0.9850
7	0.9515	0.9820	0.9840
8	0.9391	0.9859	0.9860
9	0.9291	0.9838	0.9877
10	0.9288	0.9850	0.9882
11	0.9243	0.9930	0.9957
12	0.9146	0.9970	0.9980
13	0.8940	0.9994	0.9997
14	0.8637	1.0000	1.0000

Table 3 Normalized decision matrix

Solution	Losses	SAIFI	AENS
1	0.2516	0.2925	0.2934
2	0.2524	0.2734	0.2732
3	0.2559	0.2694	0.2691
4	0.2568	0.2691	0.2690
5	0.2637	0.2659	0.2653
6	0.2642	0.2656	0.2651
7	0.2644	0.2653	0.2653
8	0.2679	0.2643	0.2648
9	0.2708	0.2649	0.2644
10	0.2709	0.2645	0.2642
11	0.2722	0.2624	0.2622
12	0.2751	0.2614	0.2616
13	0.2814	0.2607	0.2612
14	0.2913	0.2606	0.2611

Result of PROMETHEE

(See Table 8).

Distribution System Case Study 2

The application of the AHP-PROMETHEE methodology is proposed to a distribution system test network [12] based on an existing distribution network in an electricity distribution company is considered. The test network includes seven load centers, representing the accumulated load of the 11 kV distribution network at each connection point, as well as 17 existing transformers and a number of existing underground

Table 4 Weighted normalized matrix

Losses	SAIFI	AENS
0.0755	0.1024	0.1027
0.0757	0.0957	0.0956
0.0768	0.0943	0.0942
0.0770	0.0942	0.0941
0.0791	0.0931	0.0928
0.0792	0.0930	0.0928
0.0793	0.0929	0.0929
0.0804	0.0925	0.0927
0.0812	0.0927	0.0925
0.0813	0.0926	0.0925
0.0817	0.0918	0.0918
0.0825	0.0915	0.0916
0.0844	0.0913	0.0914
0.0874	0.0912	0.0914

Table 5 Best and worst

Best V+	Worst V−
0.0755	0.0874
0.0912	0.1024
0.0914	0.1027

Table 6 Separable measures

S+	S−
0.01591	0.01191
0.00620	0.01520
0.00436	0.01585
0.00435	0.01574
0.00435	0.01588
0.00438	0.01592
0.00445	0.01588
0.00523	0.01572
0.00606	0.01533
0.00605	0.01543
0.00623	0.01623
0.00706	0.01632
0.00895	0.01613
0.01191	0.01591

Table 7 Relative closeness

Pi	Ranking
0.4283	14
0.7104	10
0.7843	2
0.7835	4
0.7851	1
0.7841	3
0.7811	5
0.7502	6
0.7167	9
0.7183	8
0.7227	7
0.6980	11
0.6431	12
0.5717	13

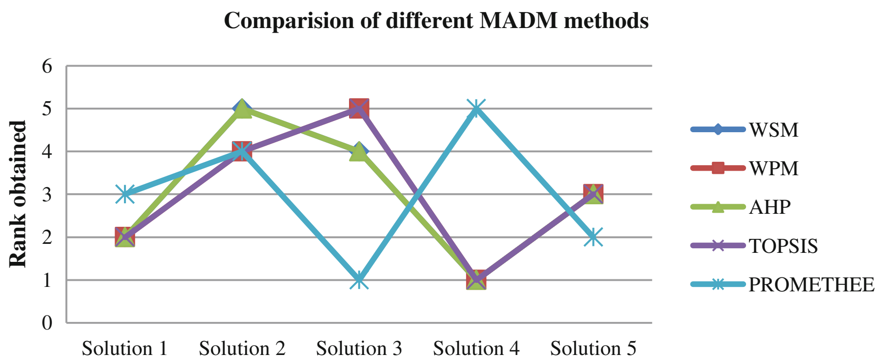


Fig. 2 Graphical representation of comparison of various MADM methods for example [12]

cables and overhead lines. In our case study, five different alternatives are to be evaluated by decision makers. Attributes shortlisted for this case study are Capital cost, Annual energy losses, System security, Supply availability, Capacity constraints, and Circuit length. All these attributes are required to be minimized as possible (Tables 9, 10) (Fig. 2).

Result of TOPSIS Method

(See Tables 9, 10, 11, 12, 13, 14, and 15).

Result of PROMETHEE

(See Table 16).

Table 8 Pi matrix, dominance, and ranking

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	Net dominance	Ranking
A1	–	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	−7.90	14
A2	0.7	–	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	−7.10	13
A3	0.7	0.7	–	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	−6.30	12
A4	0.7	0.7	0.7	–	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	−5.50	11
A5	0.7	0.7	0.7	0.7	–	0.3	0.65	0.3	0.3	0.3	0.3	0.3	0.3	0.3	−4.00	10
A6	1	1	1	1	1	–	1.65	1.3	1.3	1.3	1.3	1.3	1.3	1.3	6.60	1
A7	1	1	1	1	0.65	0.65	–	1.3	1.3	1.3	1.3	1.3	1.3	1.3	3.90	3
A8	1	1	1	1	1	1	1	–	1.65	1.65	1.3	1.3	1.3	1.3	6.10	2
A9	1	1	1	1	1	1	1	0.65	–	1.3	1.3	1.3	1.3	1.3	3.40	4
A10	1	1	1	1	1	1	1	0.65	1	–	1.3	1.3	1.3	1.3	2.80	6
A11	1	1	1	1	1	1	1	1	1	1	–	1.3	1.3	1.3	2.90	5
A12	1	1	1	1	1	1	1	1	1	1	1	–	1.3	1.3	2.30	7
A13	1	1	1	1	1	1	1	1	1	1	1	1	–	1.3	1.70	8
A14	1	1	1	1	1	1	1	1	1	1	1	1	1	–	1.10	9

Table 9 Distribution system case 2 [12]

Solution	Energy losses (MWh)	System security (% customers Interrupted)	Supply availability (CML)	Capacity constraints (MWh)	Circuit length (kM)	Capital cost (£'000)
1	14632.43	4.72	125.74	23.26	1.44	14826
2	14584.32	4.69	120.22	109.8	1.59	14813
3	14657.32	4.42	103.99	109.8	2.36	14286
4	14647.84	4.93	141.74	23.26	0.68	15703
5	14674.76	4.66	125.62	23.26	1.44	15176

Table 10 Normalized values

Solution	g1	g2	g3	g4	g5	g6
1	0.9967	0.9364	0.8270	1.0000	0.4722	0.9636
2	1.0000	0.9424	0.8650	0.2118	0.4277	0.9644
3	0.9950	1.0000	1.0000	0.2118	0.2881	1.0000
4	0.9957	0.8966	0.7337	1.0000	1.0000	0.9098
5	0.9938	0.9485	0.8278	1.0000	0.4722	0.9414

Table 11 Normalized matrix

Solution	g1	g2	g3	g4	g5	g6
1	0.4470	0.4504	0.4533	0.1450	0.4040	0.4430
2	0.4455	0.4475	0.4334	0.6844	0.4460	0.4426
3	0.4478	0.4218	0.3749	0.6844	0.6620	0.4268
4	0.4475	0.4704	0.5110	0.1450	0.1908	0.4692
5	0.4483	0.4447	0.4529	0.1450	0.4040	0.4534

Table 12 Weighted normalized matrix

Solution	g1	g2	g3	g4	g5	g6
1	0.0224	0.0676	0.0680	0.0362	0.0606	0.1107
2	0.0223	0.0671	0.0650	0.1711	0.0669	0.1106
3	0.0224	0.0633	0.0562	0.1711	0.0993	0.1067
4	0.0224	0.0706	0.0766	0.0362	0.0286	0.1173
5	0.0224	0.0667	0.0679	0.0362	0.0606	0.1134

Table 13 Ideal best and worst

Best V+	Worst V−
0.02228	0.02241
0.06326	0.07056
0.05623	0.07665
0.03625	0.17111
0.02861	0.09931
0.10671	0.11729

Table 14 Separable measures

S+	S−
0.03458	0.14076
0.14058	0.03523
0.15227	0.02412
0.02412	0.15227
0.03486	0.14069

Table 15 Performance index

Pi	Ranking
0.80279	2
0.20040	4
0.13677	5
0.86323	1
0.80140	3

Table 16 Pi matrix, dominance, and ranking

	A1	A2	A3	A4	A5	(φ^+)	(φ^-)	Net dominance	Ranking
A1	–	0.4	0.45	0.6	0.3	1.75	1.60	0.15	3
A2	0.6	–	0.2	0.6	0.45	1.85	1.90	−0.05	4
A3	0.55	0.55	–	0.55	0.6	2.25	1.50	0.75	1
A4	0.15	0.4	0.45	–	0.2	1.20	2.30	−1.10	5
A5	0.3	0.55	0.4	0.55	–	1.80	1.55	0.25	2

4 Conclusion

In this paper, MADM techniques like TOPSIS and PROMETHEE are discussed in detail and WSM, WPM, AHP, TOPSIS, and PROMETHEE are implemented for two distribution system case studies for decision-making. For case 1, the results obtained by all the MADM methods are compared and solution number 6 has obtained rank 1 in WSM, WPM, AHP, and PROMETHEE and solution number 5 has obtained rank 1 rank in TOPSIS method.

For the distribution system case 2, all the MADM methods are compared and solution number 4 has obtained rank 1 in WSM, WPM, AHP, and TOPSIS method and solution number 3 has obtained rank 1 rank by PROMETHEE.

In the TOPSIS method, the alternatives are ranked on the basis of their closeness to the best virtual alternative and distance from the worst virtual alternative. PROMETHEE is outranking method which works on pair-wise comparison of alternatives for various attributes. The methods are logical, systematic, and convenient to implement. These methods can be implemented to any decision-making problems in distribution system as decision support tool.

References

1. Shirmohammadi, D., Hong, H.W.: Reconfiguration of electric distribution networks for resistive line losses reduction. IEEE Trans. Power Syst. **4**(2), 1492–1498 (1989)
2. Kamble, S.G., Patil, U.V.: Performance improvement of distribution systems by using PROMETHEE—multiple attribute decision making method. In: International Conference on Communication and Signal Processing (ICCASP-2016)
3. Kamble, S.G., Patil, U.V., Vadirajacharya, K.: Decision making in distribution system using different MADM methods. In: National Conference on Recent trends in Electrical Engineering at Government college of Engineering, Karad
4. Kamble, S.G., Vadirajacharya, K., Patil, U.V.: Decision making in distribution systems using improved AHP-PROMETHEE method. In: IEEE International Conference on Computing Methodologies and Communication (ICCMC 2017)
5. Kamble, S.G., Vadirajacharya, K., Patil, U.V.: Application of improved TOPSIS method for decision making in distribution system. In: IEEE 2nd International Conference on Inventive Computation Technologies (ICICT 2017)
6. IEEE Guide for Electric Power Distribution Reliability Indices, IEEE Std. 1366-2003, May 2004
7. Venkata Rao, R.: Decision Making in the Manufacturing Environment, pp. 27–41. Springer London Limited (2007)
8. Venkata Rao, R., Patel, B.K.: Decision making in the manufacturing environment using an improved PROMETHEE method. Int. J. Product. Res. 1–18 (2009), iFirst
9. Venkata Rao, R.: Decision Making in the Manufacturing Environment Using Graph Theory and Fuzzy Multiple Attribute Decision Making Methods, vol. 2, Springer Series in Advanced Manufacturing, Springer London
10. Triantaphyllou, E., Shu, B., Nieto Sanchez, S., Ray, T.: Multi-criteria decision making: an operations research approach. In: Webster, J.G. (ed.) Encyclopedia of Electrical and Electronics Engineering. Wiley, New York, NY (1998)
11. Paterakis, N.G., Mazza, A., Santos, S.F., Erdinç, O., Chicco, G., Bakirtzis, A., Catalão, J.P.S.: Multi-objective reconfiguration of radial distribution systems using reliability indices. IEEE Trans. Power Syst. **31**, 1048–1062 (2016)
12. Zhang, T., Zhang, Guangquan, Jun, M.A., Lu, J.: Power distribution system planning evaluation by a fuzzy multi-criteria group decision support system. Int. J. Comput. Intell. Syst. **3**(4), 474–485 (2010)
13. Espie, P., Ault, G.W., Burt, G.M., McDonald, J.R.: Multiple criteria decision making techniques applied to electricity distribution system planning. IEE Proc.-Gener. Transm. Distrib. **150**(5), 527–535 (2003)
14. Wong, S., Bhattacharya, K., Fuller, J.D.: Electric power distribution system design and planning in a deregulated environment. IET Gener. Transm. Distrib. **3**(12), 1061–1078 (2009)
15. Mazza, A., Chicco, G., Russo, A.: Optimal multi-objective distribution system reconfiguration with multi criteria decision making-based solution ranking and enhanced genetic operators. Int. J. Elect. Power Energy Syst. **54**, 255–267 (2014)
16. Pohekar, S.D., Ramachandran, M.: Application of multi-criteria decision making to sustainable energy planning—a review. Renew. Sustain. Energy Rev. **8**, 365–381 (2004). 369
17. Vitorino, R.M., Jorge, H.M., Neves, L.P.: Multi-objective optimization using NSGA-II for power distribution system reconfiguration. Int. Trans. Electr. Energy Syst. (2013) (Wiley Online Library (wileyonlinelibrary.com)). https://doi.org/10.1002/etep.1819
18. Baran, M.E., Wu, F.F.: Network reconfiguration in distribution systems for loss reduction and load balancing. IEEE Trans. Power Deliv. **4**(2), 1492–1498 (1989)

Interconnection of Grid and Renewable Energy Sources Using Voltage Source Inverter

Anish Vijay Patil and K. Vadirajacharya

Abstract Nowadays renewable energy is used in large amount to compensate the grid power need and for onsite production in remote places. Due to variable nature of renewable energy sources and due to high penetration level of intermittent renewable energy sources they pose power quality problems to the system which can be resolved by using power electronics technology. In this paper voltage source converter based interconnection technology is demonstrated, a four leg VSI is used as interconnecting device between renewable energy source and grid network. A closed loop SPWM technique is used for controlling the VSI. It gives improved power quality features by reducing the harmonic content in the RES system and better control on duty cycle of the VSI. The performance of designed controller is verified on PSIM platform.

Keywords Renewable energy sources (RES) · Voltage source converter (VSC) Sinusoidal pulse width modulation (SPWM) · THD

1 Introduction

Conventional energy sources are being replaced with Renewable energy sources (RES) and are used as alternative energy source for fossil fuel across the globe. The renewable energy sources connected at distribution level is commonly called as Distributed Generation (DG) or Dispersed Generation. The DG connecting the renewable energy sources to grid is poses problem that the output of renewable energy sources is not a pure sinusoidal in nature. So if DG connected directly to grid, will effect on power quality [1, 2] and efficiency. Therefor for ensuring reliable and efficient grid operation semiconductor switch based voltage source converters are being used as interface between grid and renewable energy sources. With this

A. V. Patil (✉) · K. Vadirajacharya
Department of Electrical Engineering, DBATU, Lonere, India
e-mail: patilanish39@gmail.com

K. Vadirajacharya
e-mail: kvadirajacharya@dbatu.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_69

the system can be actively controlled to improve the system operation and power quality at point of common coupling. The most common and efficient renewable energy sources used in the system are solar and wind. Before connecting these non-conventional sources to grid energy conversion is obtained with the help of converters. If two or more sources are used in a system, generally they are connected to common a DC bus before grid connection for better regulation. Therefore in case of wind solar hybrid system, wind mills are interfaced with AC-to-DC converter, while PV solar are interfaced with DC-to-DC converters with stiff DC bus. The stiff DC bus supplies power from DC bus to the power grid using Voltage Source Inverter [3] generally the output of Voltage Source Inverters which are used for converting DC power to AC power are controlled using different control strategies. Commonly used control strategies for controlling voltage source inverter are pulse width modulation, DSP-based controller, Hysteresis current controller etc. In this paper a sine wave pulse width modulation technique is being used for controlling output voltage of voltage source inverter. In order to compensate neutral current, the voltage source inverter used is a 4 leg three phase four wire system. The voltage source inverter with selected configuration and controlling method provides required power to the grid at rated current, voltage and frequency at acceptable level of power quality of the system [4, 5].

2 Operating Principle

2.1 DC Link Operation

The output of the renewable energy sources is not constant and poses problem while connecting to grid. This problem of variable output of renewable energy sources is over come with help of DC link. The DC link regulates power fed to grid and provides continues power output even if the input is variable in nature.

The current output of DC link is given by.

$$I_{dc} = \frac{P_{res}}{V_{dc}}, \tag{1}$$

where P_{res} = Power supplied by renewable energy source,

The current flow at the output side of the DC link is given by,

$$I_{dc2} = \frac{P_{inv}}{V_{dc}} = \frac{P_g + P_{loss}}{V_{dc}}, \tag{2}$$

where

P_{inv} = Total power output of inverter at grid interface,
P_g = Active Power of grid,

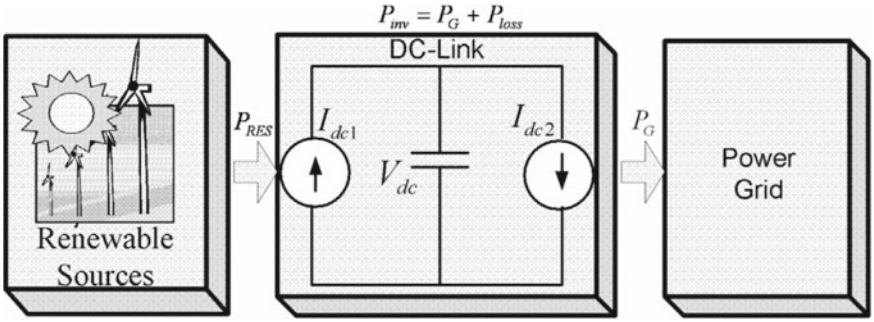


Fig. 1 DC Link equivalent diagram

P_{loss} = Power loss in inverter,

If losses are neglected then we can write (Fig. 1),

$$P_{res} = P_g$$
$$P_{inv} = P_g + P_{loss}$$

2.2 Control of Voltage Source Inverter

The voltage source inverter used is controlled in such a way that it will ether absorb or supplies active power from or to grid. Control approach is designed in such a way that it will compensate harmonics, unbalance and change in neutral current which is introduced in the system due to nonlinear or unbalance load connected at point of common coupling. The reference current for control strategy is obtained by multiplying active current and unit grid voltage vector, the unit voltage vector template can be given as.

$$V_a = (sin\,\theta) \tag{3}$$

$$V_b = \left(sin\,\theta - \frac{2\pi}{3}\right) \tag{4}$$

$$V_c = \left(sin\,\theta + \frac{2\pi}{3}\right) \tag{5}$$

The error obtain in the DC link voltage is indicated as V_{dcerr}, at N-th sampling instant the error is given as,

$$V_{dcerr(n)} = V^*_{dc(n)} - V_{dc(n)} \tag{6}$$

At the N-th instant output of PI controller is given as,

Table 1 System parameter

Name	Value
3-phase supply	230 V, 50 Hz
3-phase load	22.66 Ω, 10 mH
DC link capacitance and voltage	3000 μF, 90 V
Coupling inductance	2.0 mH

$$I_m = I_{m(n-1)} + K_{pv_{dc}}\left(V_{dcerr(n)} - V_{dcerr(n-1)}\right) + K_{iv_{dc}} V_{dcerr(n)}, \tag{7}$$

where $K_{pvdc} = 10$ and $K_{ivdc} = 0.05$,

This are the proportional and integral gain of PI controller.

As mentioned above the value of reference field grid current is given as,

$$I_a^* = I_m \cdot V_a \tag{8}$$

$$I_b^* = I_m \cdot V_b \tag{9}$$

$$I_c^* = I_m \cdot V_c \tag{10}$$

$$I_n^* = 0 \tag{11}$$

For computing the current errors the reference grid currents are compared with actual grid current as,

$$I_{aerr} = I_a^* - I_a \tag{12}$$

$$I_{berr} = I_b^* - I_b \tag{13}$$

$$I_{cerr} = I_c^* - I_c \tag{14}$$

3 Simulation and Results

3.1 Simulation

The proposed system consists of DC source representing a stiff solar PV module and wind mill connected to dc link. The dc link output is given to the four leg inverter output is then connected to the system grid. The control method used for controlling the VSI is a simple SPWM technique, along with a close loop system which is used for controlling the duty cycle of the VSI with reference to the system voltage and load current. The system parameters indicated below are taken from the reference [1] and simulated on PSIM platform for performance verification (Table 1) (Fig. 2).

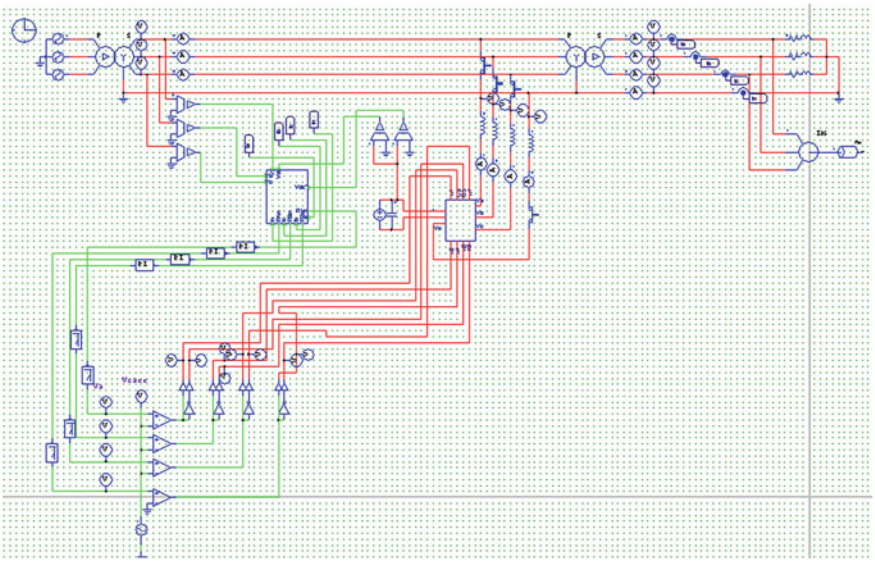


Fig. 2 System on simulation

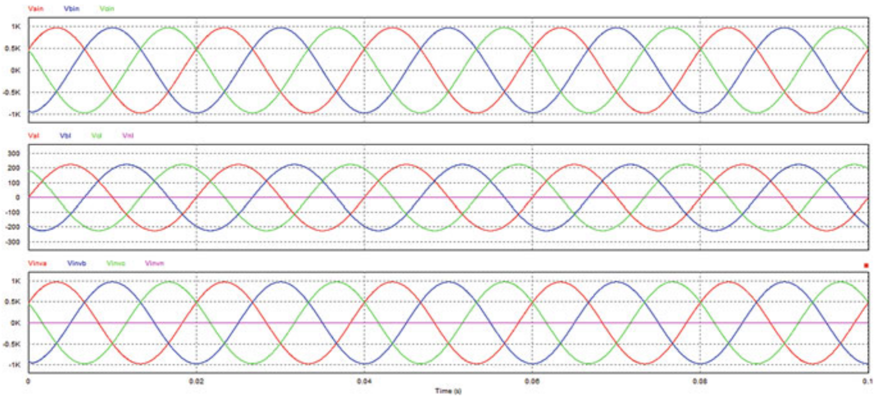


Fig. 3 Grid voltage, load voltage and inverter voltage waveform

4 Observation and Analysis

Figure 3 shows waveform of the grid voltage, load voltage, and inverter voltage respectively for a linear RL load.

Figure 4 shows waveform of the grid current, inverter current and load current respectively. From the waveforms it is clear that the grid and RES are equally sharing the load that means the impedance of grid and RES is equal. It is also clear that the grid current and inverter current follows the load current and the inverter current

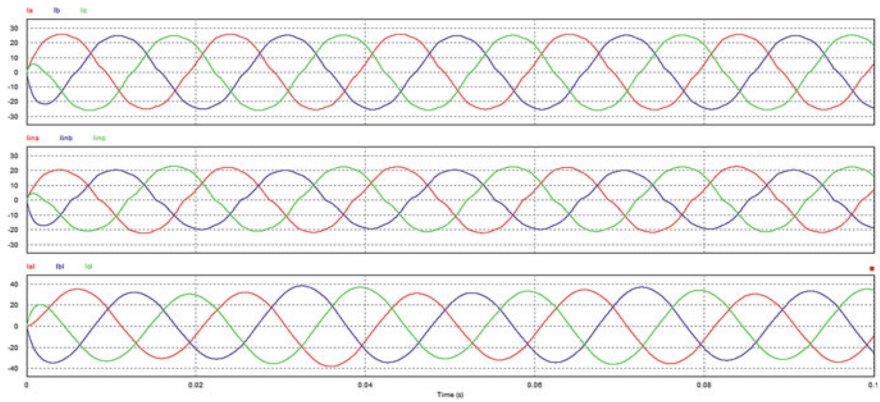


Fig. 4 Grid current, inverter current and load current waveform

lags the inverter voltage due to the delay angle of the inverter. The performance of the designed system is verified for a linear RL load which dissipates satisfactory performance.

5 Conclusion

Considering demerits of conventional sources and merits of renewable energy sources RES are being used as supporting energy sources in electric power systems along with conventional power systems. The renewable energy sources are intermittent in nature. The grid requires constant sources. Thus from above results it is seen that while connecting intermittent sources like solar and wind to the grid if connected through interface converter through stiff DC link can provide satisfactory operation for the grid integration. The output of Voltage Source Inverter is controlled using a Sinusoidal Pulse Width Modulation Technique. Close loop system controls the duty cycle of the inverter by considering grid voltage and load current which gives real time controlling of the inverter.

References

1. Singh, M., Khadkikar, V., Chandra, A., Varma, R.K.: Grid interconnection of renewable energy sources at the distribution level with power-quality improvement features. IEEE Trans. Power Deliv. **26**(1) (2011)
2. Nibitha, N.S., Soumya, A.V.: The analysis of grid interconnected system at distribution level using renewable energy resources. Int. J. Technol. Enhanc. Emerg. Eng. Res. (IJTEEE) **3**(10), 53–58 (2015). ISSN 2347-4289

3. Trip, N.D., Şchiop, A.: Considerations on efficiency increase for power electronic circuits that interconnect renewable energy sources to the grid. In: Proceedings of ISETC 2016 Timisoara, Romania, 27–28 Oct 2016
4. Kantar, E., Hava, A.M.: Design of grid connected PWM converters considering topology and PWM methods for low voltage renewable energy applications. In: Proceedings of IPEC Hiroshima, Japan, 2014, 18–21 May 2014
5. Trip, N.D., Neamţu, M.O.: DSP based interconnection circuit of the renewable energy sources to a smart grid. In: Proceedings of SIITME 2016 IEEE 22nd International Symposium for Design and Technology in Electronic Packaging, Oradea, Romania, 20–23 Oct 2016
6. Shitole, B., Suryawanshi, H.M.: Comparative evaluation of synchronization techniques for grid interconnection of renewable energy sources. In: Proceedings of IECON 2015, 9–12 Nov 2015
7. Wang, Y., Zhao, J., Dai, J., Liu, H., Ushiki, S., Ohshima, M.: Evaluation and testing of voltage-controlled grid-connected inverter for high penetration of renewable energy. In: Proceedings of APEC 2015 Charlotte, NC, USA, 15–19 Mar 2015
8. Vinifa, R., Kavitha, A.: Linear quadratic regulator based current control of grid connected inverter for renewable energy applications. In: Proceedings of ICEETS 2016 India, 7–8 Apr 2016

Performance Comparison of Sliding Control Law for Dynamical Systems

S. S. Sankeswari and R. H. Chile

Abstract In this paper, a sliding mode control approach with PID sliding surface and first-order filter are implemented for single-input-single-output (SISO) devices. The proposed controller performs satisfactorily even for parameter variations in the system. In order to verify the applicability to the disturbances, an external load is applied and the performance of the controller is validated. The stability can be tested using the concept of Lyapunov stability theorem. In sliding mode approach, the effect of chattering phenomenon is significantly reduced by selecting appropriate switching (gain) of interest along with the known parameters of the system. In case of the proposed approach, it can verified that this approach has applications to the physical systems. The applicability and the performance of the proposed control structure are confirmed by a simulation example and the analysis of the proposed controller is carried out with the similar controllers available in the literature.

Keywords Sliding mode control · Single variable · Stability · Robustness

1 Introduction

In practical industrial applications, almost all the systems have nonlinear behavior in nature, but for the purpose of analysis and control, it has been approximated to linear model. The control of these systems which are affected by parameter inaccuracy, i.e., matched uncertainty, unmodelled dynamics with time delay, which is generally approximated have been a serious challenge to the controller design [1].

S. S. Sankeswari (✉)
Department of Electrical Engineering, Mahatma Bashveshwar Education Society's
College of Engineering, Ambajogai, India
e-mail: sankeswari@gmail.com

R. H. Chile
Department of Electrical Engineering, SGGS Institute of Engineering and Technology,
Nanded 431606, India
e-mail: rhchile@yahoo.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_70

The sliding mode controller and feedback linearization concepts of the controller design techniques for nonlinear systems have been promising and useful in control problems, where the approximate linear model of the systems are available [2, 3]. In early 1950s, the theory of sliding mode control was evolved and it can be applied effectively to tackle the systems or processes with uncertainties in addition to external disturbances [4].

The main contribution and motivation for the proposed work are that the desired performance of system with or without external uncertainties can be achieved by means of selecting an appropriate and simple choice of the sliding surface. The sliding mode controller has a wide variety of applications and the related works are available in the literature, for example, for gyroscope application [5], robot applications [6], and turbulent flows application [7]. Tannuri et al. [8] and Lee et al. [9] designed the controller for position control system and the higher order systems are controlled by designing conventional controller [10] and sliding mode controller [11]. The work of Eker is mostly for the application of electromechanical speed control using conventional sliding or second-order sliding mode control [12, 13]. Recently, Furat and Eker reported that the speed of the electromechanical system can be controlled with an experimental application of second-order integral sliding mode control [14].

The outline of the paper is as follows. In the Sect. 2, the descriptions of the typical second-order models are included. In Sect. 3, sliding surface design is presented and the total control law is derived, while Sect. 4 give a typical example and its closed-loop performance comparisons to verify the control capability in addition to effectiveness of the proposed controller. In Sect. 5, concluding remarks as well as future direction of the work is included.

2 Descriptions of Systems

The typical second-order systems is given by

$$\frac{Y(s)}{U(s)} = \frac{K\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} = \frac{C_n}{s^2 + A_n s + B_n} \tag{1}$$

where K is gain of the system, ζ and ω_n are damping factor and natural frequency of the systems, respectively. Any system of interest needs to be converted into second-order system as given in Eq. 1.

2.1 First-Order Plus Delay Time Systems

The FOPDT systems are modeled in the structure as follows:

$$\frac{Y(s)}{U(s)} = \frac{Ke^{-t_d s}}{\tau s + 1} \tag{2}$$

where the notation τ is a time constant, t_d is a time delay, and K is a steady-state gain. For smaller time delay, as compared to the time constant τ, the system in the above Eq. (2) can be approximated as below [15]

$$\frac{Y(s)}{U(s)} = \frac{K}{(\tau s + 1)(t_d s + 1)} = \frac{C_n}{s^2 + A_n s + B_n} \tag{3}$$

In this case, the approximation of time delay $e^{-t_d s} = 1/(t_d s + 1)$ is called Taylor Series approximation. It is a common practice in controller design process to use the Taylor series approximation for time delay system [16].

3 Sliding Mode Control Approach

Consider a sliding surface of PID type with first-order filter, as follows:

$$\sigma(t) = \left[K_p + \frac{K_i}{s} + sK_d\right]^{n-1} \psi(E(s)) \tag{4}$$

where $\psi(E(s))$ is the filter and it is given as, $\psi(E(s)) = 1/(\lambda s+1)E(s)$ and "n" is the system order. The tuning parameters K_p, K_i, K_d, and λ which helps for designing the sliding surface $\sigma(t)$ and is a choice of control designer that defines the system performance during the sliding phase. Here, the objective is to design the controller which ensures that the plant output $y(t)$ should track the reference input $r(t)$ as applied. This means that the error signal and its respective derivatives should converge to zero. The intention of designing sliding mode control is to make ensure that the error e(t) will converge to zero in accordance with the sliding surface. Substituting $\psi(E(s)) = 1/(\lambda s + 1)E(s)$ in Eq. 4, gives

$$\sigma(t) = \frac{1}{\lambda s + 1} K_p E(s) + \frac{K_i}{s(\lambda s + 1)} E(s) + \frac{1}{\lambda s + 1} K_d s E(s) \tag{5}$$

The second-order transfer function means $n = 2$.

The error dynamics in second order are

$$\ddot{e}(t) = \ddot{r}(t) - \ddot{y}(t) \tag{6}$$

where reference input is r(t) and plant output is y(t).

In general, from Eq. (3), $\ddot{y}(t) = -A_n\dot{y}(t) - B_n y(t) + C_n u(t) + D(t, u(t))$.

Put $\ddot{y}(t) = -A_n\dot{y}(t) - B_n y(t) + C_n u(t) + D(t, u(t))$ in Eq. (6)

$$\ddot{e}(t) = \ddot{r}(t) - [-A_n\dot{y}(t) - B_n y(t) + C_n u(t) + D(t, u(t))] \tag{7}$$

$$\ddot{e}(t) = \ddot{r}(t) + A_n\dot{y}(t) + B_n y(t) - C_n u(t) - D(t, u(t)) \tag{8}$$

The second-order derivative in Eq. (5) is obtained by multiplying both sides by "$s(\lambda s + 1)$". Thus, Eq. (5) can be rewritten as

$$s(\lambda s + 1)\sigma(t) = sK_p E(s) + K_i E(s) + s^2 K_d E(s) \tag{9}$$

After converting the above Eq. (9) into time domain, it can be written as

$$\ddot{\sigma}(t) = \frac{K_p}{\lambda}\dot{e}(t) + \frac{K_i}{\lambda}e(t) + \frac{K_d}{\lambda}\ddot{e}(t) - \frac{\sigma(t)}{\lambda} \tag{10}$$

From Eq. (7), we know, $\ddot{e}(t) = \ddot{r}(t) - [-A_n\dot{y}(t) - B_n y(t) + C_n u(t) + D(t, u(t))]$. Using this relation Eq. (10), can be modified as

$$\ddot{\sigma}(t) = \frac{K_p}{\lambda}\dot{e}(t) + \frac{K_i}{\lambda}e(t) + \frac{K_d}{\lambda}\begin{bmatrix} \ddot{r}(t) + A_n\dot{y}(t) + B_n y(t) \\ -C_n u(t) - D(t, u(t)) \end{bmatrix} - \frac{\dot{\sigma}(t)}{\lambda} \tag{11}$$

If the conditions are like $\sigma(t) = \dot{\sigma}(t)$ and $\ddot{\sigma}(t) = 0$ using the equivalent control $u(t) = u_{eq}(t)$ the designed SMC is called as a second-order SMC. The equivalent control is designed with the help of system which working under nominal conditions, i.e., $D(t, u(t)) = 0$, as follows:

substituting $\ddot{\sigma}(t) = 0$, in Eq. (11)

$$\frac{K_p}{\lambda}\dot{e}(t) + \frac{K_i}{\lambda}e(t) + \frac{K_d}{\lambda}\begin{bmatrix} \ddot{r}(t) + A_n\dot{y}(t) + B_n y(t) \\ -C_n u(t) - D(t, u(t)) \end{bmatrix} - \frac{\dot{\sigma}(t)}{\lambda} = 0 \tag{12}$$

Let $u = u_{eq}$ from the above Eq. (12)

$$\begin{aligned} u_{eq}(t) = {} & \frac{1}{(K_d C_n)}\left(K_p\dot{e}(t) + K_i e(t) + K_d\ddot{r}(t) + K_d A_n\dot{y}(t) + K_d B_n y(t)\right) \\ & + \frac{1}{(K_d C_n)}\left(\frac{-K_p}{\lambda}e^{(-t/\lambda)}e(t) - K_i e(t) + K_i e^{(-t/\lambda)}e(t) - \frac{K_d}{\lambda}e^{(-t/\lambda)}\dot{e}(t)\right) \end{aligned} \tag{13}$$

The above Eq. 13 is called equivalent controller. The input in SMC can be split as

$$u(t) = u_{eq}(t) + u_{sw}(t) \tag{14}$$

Now, The switching control taken in this work is as follows:

$$u_{sw}(t) = k_{sw}r^2(t)\tilde{e}(t)\mathrm{sgn}\left(\frac{k_{sf}}{\tilde{e}(t)}\dot{\sigma}(t)\right) + \frac{1}{K_d C_n}\mathrm{sgn}(\sigma(t))s \tag{15}$$

where k_{sw} is a switching gain which is positive and used to limit the chattering phenomenon while maintaining the tracking performance by assuming $r(t) \neq 0$ and defining $\tilde{e}(t)$ as a modified error, defined as follows:

$$\tilde{e}(t) = \in_1 \operatorname{sgn}(e(t)) \ldots if \ldots |e(t)| \leq \in_1 \tag{16}$$

$$\tilde{e}(t) = e(t) \ldots if \ldots |e(t)| \geq \in_1 \tag{17}$$

where $\in > 0$ is a small number which avoids the issue of zero division.

3.1 Stability Condition

In case of control system design, the stability and robustness issues need to be addressed for applicability of the method. The stability of the proposed controller is guaranteed through Lyapunov stability criteria. The choice of Lyapunov function is given as follows:

$$V(t) = \frac{1}{2}\dot{\sigma}^2(t) + |\sigma(t)| \tag{18}$$

The function in Eq. (18) is satisfied for all $\sigma(t)$ not equal to zero. By taking derivative, the Eq. (18) can be written as

$$\dot{V}(t) = \dot{\sigma}(t)\ddot{\sigma}(t) + \dot{\sigma}(t)\frac{|\sigma(t)|}{\sigma(t)} \tag{19}$$

or

$$\dot{V}(t) = \dot{\sigma}(t)\left[\ddot{\sigma}(t) + \frac{|\sigma(t)|}{\sigma(t)}\right] \tag{20}$$

As time t$\rightarrow \infty$ the output y(t)$\rightarrow$r(t) and therefore after settling time the error signal $e(t) \approx 0$. With $e(t) = 0$, $\dot{V}(t)$ can be rewritten as

$$\dot{V}(t) = \frac{\dot{\sigma}(t)}{\lambda}[S_1] \leq 0 \tag{21}$$

where $S_1 = K_d C_n k_{sw} r^2 + \widetilde{e}(t) \mathrm{sgm}\left(\frac{k_{sf}}{\widetilde{e}(t)}\right)\dot{\sigma}(t) + \mathrm{sgm}(\sigma(t)) + K_d C_n n - K_d D(t, u(t))$

Therefore, the condition for the stability is

$$k_{sw} > \frac{D\max}{C_n r^2(t)\left|\tilde{e(t)}\right|} \tag{22}$$

An appropriate selection of the value k_{sw} is necessary because too low value leads to very slow convergence and large value causes chattering. This bound on the parameters k_{sw} ensure system stability. By keeping the condition on the tuning parameter, the same method can be extended to higher order systems.

4 Simulation Example

A brushless DC Motor (BLDC) is used in simulation to obtain the results of sliding mode control. The results of simulation justify the effectiveness of the proposed controller and the results of method proposed are compared with sliding mode controllers available in the literature. The BLDC motor presented in this paper is the EC 45 flat Ø 45 mm, brushless, 30 W from Maxon motors [17]. The parameters given in the datasheet of Maxon is used to obtain transfer function of BLDC motor given as.

$$G(s) = \frac{1/K_e}{\tau_m \tau_e s^2 + \tau_m s + 1}$$

where K_g, τ_m and τ_e are the constants and need to calculated.

The term τ_e is calculated using

$$\tau_e = \frac{L}{3R} = \frac{0.560 \times 10^{-3}}{3 \times 1.10} = 151.56 \times 10^{-6}$$

The term τ_m is calculated using

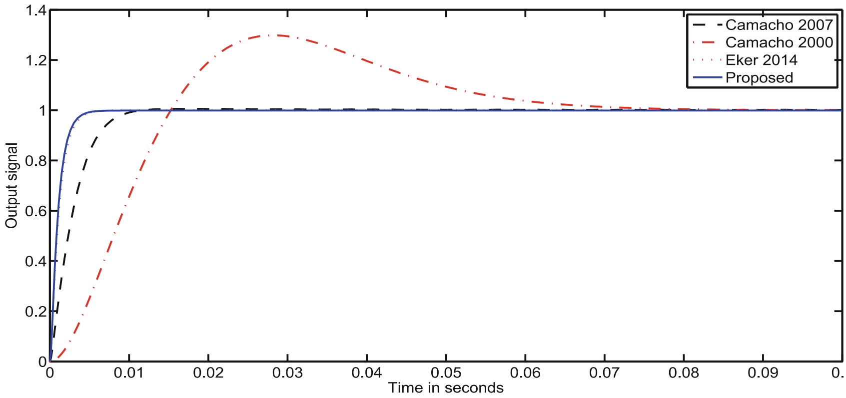
$$\tau_m = \frac{3R_{\emptyset} J}{K_g K_t} = 0.0171 \quad \text{with} \quad K_e = \frac{3R_{\emptyset} J}{\tau_m K_t} = 0.0763$$

Thus, the model of the DC motor in the form of transfer function is

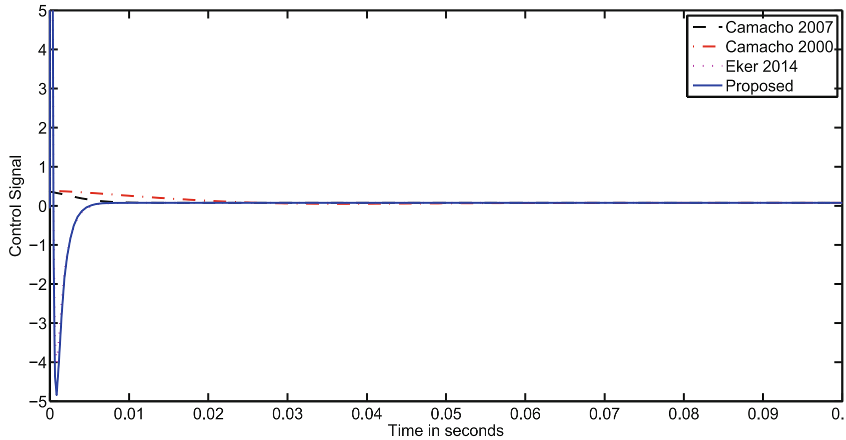
$$G(s) = \frac{13.11}{155.56 \times 0.0171 \times 10^{-6} s^2 + 0.0171 s + 1} = \frac{82620}{s^2 + 269.7s + 6302}$$

The parameters used for the proposed and Furat and Eker [14] are

$k_{sw} = 200; k_{sf} = 0.025; K_p = 12; K_i = 0.001; K_d = 0.0024$; For the proposed method filter parameter $\lambda = 0.9$ is selected, the performance of the proposed SMC and other controllers in terms of output responses and input responses are shown in Fig. 1a and Fig. 1b, respectively. It is clear from the output responses that the controller given by Furat and Eker and proposed controllers gives fast and satisfactory response while the responses of Camacho-2000 and Camacho-2007 are sluggish. The response given by Camacho-2000 gives large overshoot and is not suitable for applications like speed control of DC motor. The control signal given by proposed controller is smooth and sliding surface is stable.



(a) Closed- loop output responses



(b) Input responses(Controller action)

Fig. 1 Simulation results of Maxon DC motor

In order to verify the robustness of the controller and its performance under external disturbances, the output disturbance $d = 0.2r$ is added at time $t = 0.05$ s. The output responses by all controllers are shown in Fig. 2. It is obvious from Fig. 2 that the controller given by Camacho-2000 and Camacho-2007 cannot be used as its output responses are not satisfactory. The proposed controller gives comparative and acceptable performance. From the discussion so far with and without external disturbances, it clear that the proposed method provides suitable solution in terms of peak overshoot and settling time in the responses in qualitative view.

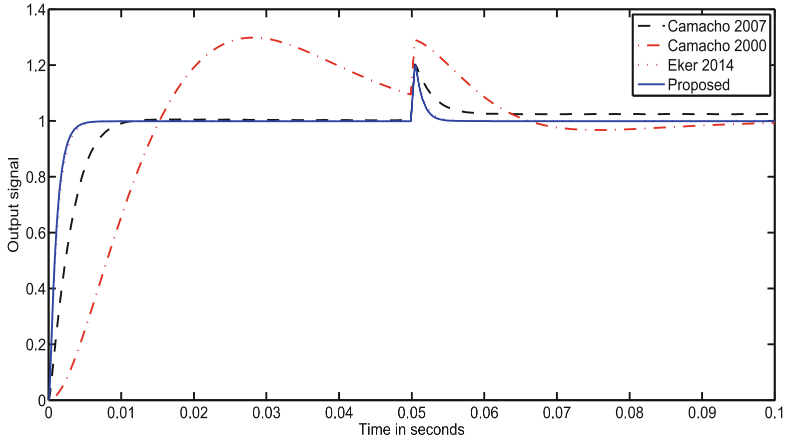


Fig. 2 Output responses under 20% external disturbances

5 Conclusion

From the simulation results, it is observed that the proposed sliding mode controller exhibits better performance. From the output response, it is observed that, no overshoot, minimum settling time, and smaller rise time in magnitude were obtained from the proposed control system. The SMC of Camacho [2000] [18] did not meet the needs of precise control, as it resulted in large overshoots and in high magnitude of settling times. The proposed SMC provides important advantages over the others like limiting the overshoot in speed, an improvement in the settling time and reducing the rise time. The simulation results of Camacho [2007], SMC is acceptable when the system has nominal parameters while it is not acceptable in case of system with parametric uncertainties. It can be concluded that the proposed SMC is comparable with Furat and Eker with simple design procedure and can be used for the system where uncertain parameters cannot be judged or measured. The proposed controller would be useful in case of external disturbances. The work can be extended for the systems with greater than ±20% parametric uncertainties by modifying the control law. This study can be extended to real-time applications by designing an appropriate experimental set-up and interfacing accessories for DC drives.

References

1. Slotine, J.J., Li, W.: Applied Nonlinear Control. Prentice-Hall, Englewood Cliffs, NJ (1991)
2. Utkin, V.I.: Sliding Modes and Their Applications in Variable Structure Systems. MIR Publishers, Moscow, USSR (1978)
3. Utkin, V.I.: Sliding mode control design principles and applications to electric drives. IEEE Trans. Ind. Electron. **40**, 23–26 (1993)

4. Liang, C.H., Su, J.P.: A new approach to the design of a fuzzy sliding mode controller. Fuzzy Sets Syst. **139**, 111–124 (2003)
5. Zhang, R., Shao, T., Zhao, W., Li, A., Xu, B.: Sliding mode control of MEMS gyroscopes using composite learning. Neurocomputing (2017) (In press, accepted manuscript)
6. Zhao, Y., Huang, P., Zhang, F.: Dynamic modeling and super twisting sliding mode control for tethered space robot. Acta Astronaut. (2017) (In press, accepted manuscript)
7. Feingesicht, M., Polyakov, A., Kerherv, F., Richard, J.-P.: Sliding mode control for turbulent flows. IFAC-PapersOnLine **50**(1), 2690–2695 (2017)
8. Tannuri, E.A., Agostinho, A.C., Morishita, H.M., Moratelli, L. Jr.: Dynamic positioning systems: an experimental analysis of sliding mode control. Control Eng. Pract. (2010). https://doi.org/10.1016/j.conengprac.2010.06.007
9. Lee, W.R., Lee, J.H., You, K.H.: Augmented sliding-mode control of an ultra precision positioning system. Precis. Eng. **35**, 521–524 (2011)
10. Malwatkar, G.M., Sonawane, S.H., Waghmare, L.M.: Tuning PID controllers for higher-order oscillatory systems with improved performance. ISA Trans. **48**, 347–353 (2009)
11. Lu, K., Xia, Y., Zhu, Z., Basin, M.V.: Sliding mode attitude tracking of rigid spacecraft with disturbances. J. Frankl. Inst. (2011). https://doi.org/10.1016/j.jfranklin.2011.07.019
12. Eker, I.: Second-order sliding mode control with experimental application. ISA Trans. **49**, 394–405 (2010)
13. Eker, I.: Sliding mode control with PID sliding surface and experimental application to an electromechanical plant. ISA Trans. **45**, 109–118 (2006)
14. Furat, M., Eker, I.: Second-order integral sliding mode control with experimental application. ISA Trans. (2014)
15. Camacho, O., Rojas, R., Gabin, W.G.: Some long time delay sliding mode control approaches. ISA Trans. **46**, 95–101 (2007)
16. Utkin, V.I.: Sliding Modes in Control and Optimization. Springer, Heidelberg, Berlin (1992)
17. Maxon Ec motor, May 2008 Edition, EC 45 at 45 mm, brushless, 30 watt maxon at motor
18. Camacho, O., Smith, C.A.: Sliding mode control: an approach to regulate nonlinear chemical processes. ISA Trans. **39**, 205–218 (2000)

A Novel Method for Detection of Atrial Fibrillation Based on Heart Rate Variability

Akib Shah and Vaishali Ingale

Abstract Atrial Fibrillation (AF) detection is one of most important part of clinical testing. We propose a novel method to detect AF episodes based on heart rate variability feature of AF. In this method, scatter plot of the heart rate is found and histogram of the Y axis data of scatter plot is used for calculations. Depending upon amount of data present in each bin of histogram ECG signal is classified as Atrial Fibrillation (AF) or Normal Sinus Rhythm (NSR). Physionet 2017 challenge database, MIT-BIH AF database and MIT-BIH NSR database are used to validate the algorithm. Physionet Challenge contains 5787 ECG records of 30/60 s classified as AF or NSR, MIT-BIH AF database contains 25 full length ECG records and MIT-BIH NSR database contains 18 full length ECG records. Using the method, we got the accuracy of 97.23% for Physionet 2017 challenge database and 97.15% for MIT-BIH AF database. MIT-BIH NSR database didn't show any AF episode. This method can also be used for real time monitoring of ECG for AF detection.

Keywords Atrial fibrillation · Heart rate variability · Scatter plot · Histogram

1 Introduction

One the most common type of arrhythmia is Atrial Fibrillation (AF). Heart rate is increased in AF, accompanied by irregular heartbeats. AF can cause strokes and hypertension [1] resulting in decrease in life expectancy of people [2]. AF detection can be difficult as it occurs only in few episodes of ECG. Many ways have been found to detect the AF episodes in the ECG. These methods can be divided into two main categories, namely atrial activity and ventricular response. Atrial activity includes

A. Shah (✉) · V. Ingale
Department of Electronics and Telecommunications, College of Engineering, Pune (COEP), Pune, India
e-mail: akibshah7897@gmail.com; shahas15.extc@coep.ac.in

V. Ingale
e-mail: vvi.extc@coep.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_71

detection of P-wave absence [3], f wave presence [4] and wavelet entropy [5]. P-wave absence detection can be improved using training set and neural network [6]. These methods become inefficient when noise is introduced in the ECG. Ventricular response consists of heart rate variability [7]. This method is preferable in comparison to the atrial activity based methods since it is more immune to the noise in the ECG. Some of heart rate variability methods are based on Poincare plot based AF detection [8], ΔRR and RR interval based AF detection [9]. Some algorithms are combination of two methods, like wavelet transform with heart rate variability [10]. This further increases the efficiency of the algorithm.

2 Methodology

The first and foremost task is noise and baseline wandering removal. Baseline wandering removal is done using the wavelet decomposition. ECG signal is decomposed up to the 10th level using the 'db6' wavelet. Then the data is reconstructed using the approximate coefficients of decomposed signal. This reconstructed signal is subtracted from the original signal to remove baseline wandering. For removing the high frequency noise, the data was passed through a low pass filter of cut-off frequency 50 Hz. This removes the fast changes from the signal thus making data smoother.

A comparative study was done between the adaptive and non-adaptive thresholding algorithms to find the R peaks of the ECG signal. For non-adaptive thresholding a simple fixed threshold was applied.

$$threshold level = \frac{\text{max\_val\_in\_data} - \text{mean\_of\_data}}{2}$$

$$\text{max\_val\_in\_data} = \text{maximum value in whole data}$$

$$\text{mean\_of\_data} = \text{mean of the whole data}$$

For non-adaptive thresholding Pan-Tompkins algorithm was chosen [11]. We found out that non-adaptive thresholding technique provides accuracy of 78% while Pan-Tompkins algorithm provides accuracy of 98.75%. Further, if the signal contains muscle artefacts the accuracy of non-adaptive thresholding algorithm decreased, while the Pan-Tompkins algorithm's accuracy was not affected much. Due to this Pan-Tompkins algorithm is used in further work.

After getting the R locations, we calculated the heart rate using consecutive RR interval. This heart rate data was used to plot the scatter plot for heart rate [i + 1] versus heart rate [i]. Using scatter plot, we tried to identify whether patient has AF episode or not. Figures 1 and 2 shows the scatter plot for AF data along with its histogram of data on Y axis. Figures 3 and 4 show the scatter plot for Normal Sinus

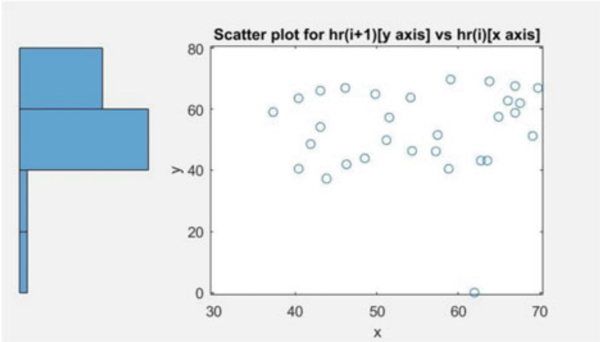


Fig. 1 Scatter plot for AF data

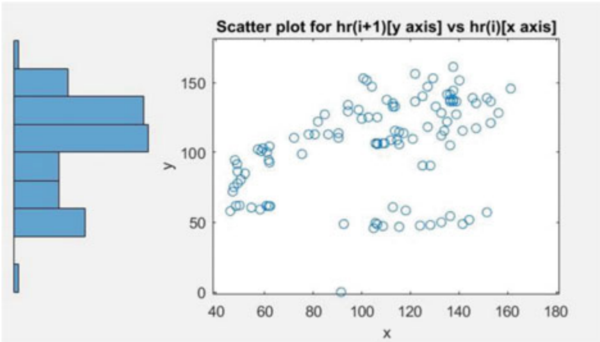


Fig. 2 Scatter plot for AF data

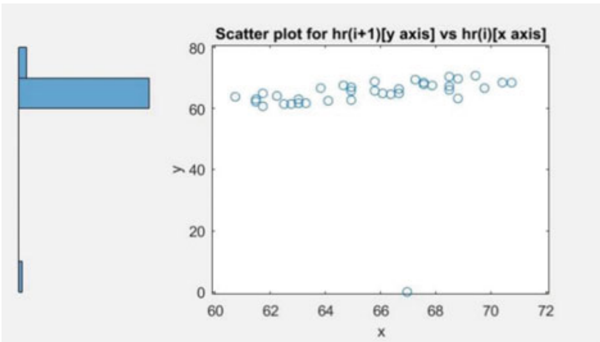


Fig. 3 Scatter plot for NSR data

Rhythm (NSR) data along with its histogram of data on Y axis. NSR means patient do not have any arrhythmia.

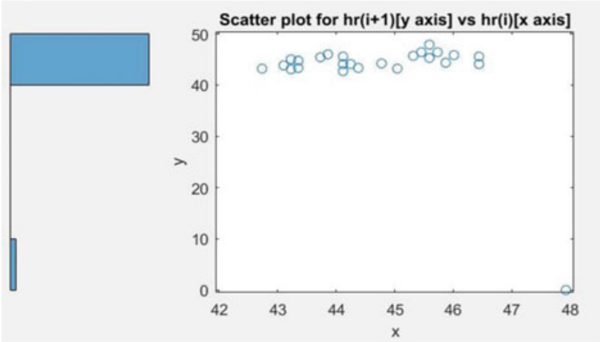


Fig. 4 Scatter plot for NSR data

2.1 Method 1

Comparing plots from Figs. 1, 2, 3 and 4, in NSR scatter plot most of the data is concentrated in a single bin of the histogram, while in AF scatter plot the data is spread over more than one bin. We used this bin counts and number of bins to determine spreading of data. This is just like finding variance of the data using bin count of histogram. Number of bins can vary according to the length of the ECG signal. By analyzing large data set of various ECG waveforms, a threshold is defined for the number of bins to be 7. Depending upon the number of bins, a logic is developed to classify the input ECG signal as AF or NSR.

If the number of bins is less than 7, then the condition for ECG being NSR is as follows: 90% of the data on Y axis of scatter plot should be present in only one bin of the histogram. If this condition is not satisfied, then the ECG is of type AF.

If the number of bins is more than 7, then the condition for ECG being a NSR is as follows: 90% of the data on Y axis of scatter plot should be present in at the most two adjacent bins of histogram. If this condition is not satisfied, then the ECG is of type AF.

2.2 Method 2

Apart from this, difference between two adjacent heart rates is also plotted. In this case more than 50% data on Y axis should be between −5 and 5. This range is found out by analyzing large dataset of ECG. This was also used to predict whether the ECG is NSR or AF.

A fuzzy logic is developed which takes input from above two methods and predicts if the signal is AF or NSR. The logic is as follows. When method 1 and method 2 both gives output as AF then the signal is assigned as AF signal. When both methods give output as NSR then the signal is assigned as NSR signal. When both methods

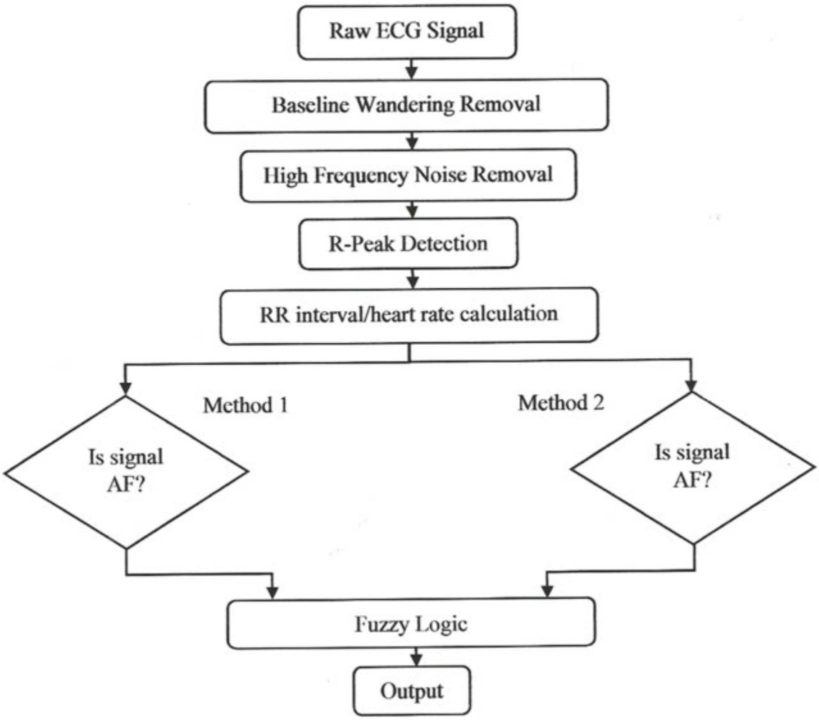


Fig. 5 Flow chart for the whole process

give different output then, from method 1 the percentage of data in most dense bin is used to classify the signal as AF or NSR. In this case, if the data, that is number of points on Y axis, in most dense bin is less than 60% then the signal is AF else, it is NSR (Fig. 5).

3 Results

We used MIT-BIH AF database, MIT-BIH NSR database and Physionet 2017 Challenge database for this study [12]. MIT-BIH AF database contains 25 full length ECG. Annotations for the onset/offset of AF episode in these ECG recordings are given by experts. MIT-BIH NSR database contains 18 long-term ECG recordings. This database does not contain any arrhythmia and hence doesn't contain any AF episodes. Physionet's training set contains 8528 ECG recordings of 30/60 s length. These recordings are divided into four categories: AF, NSR, Other and Noisy. In this study we have used recordings of AF and NSR type only (Table 1).

Table 1 Physionet 2017 challenge database

Type of record	Total records	Error output
NSR	5049	127
AF	738	33

Table 2 MIT-BIH AF database

Type of episode	Total episode present	Detected episodes
NSR	298	291
AF	299	289

Physionet database for 2017 challenge consists of 5049 NSR signals and 738 AF signals. 127 out of 5049 gave false positive outputs while 33 of 738 gave false negative output. Thus, the method gave us the accuracy of 97.23% for this database.

For MIT-BIH AF database, 25 records contain total 299 AF episodes and 298 NSR episodes which are shown in Table 2. Of 299 AF episodes 289 were detected, while of 298 NSR episodes 291 were detected. This gave us the sensitivity of 96.65%, specificity of 97.65% and accuracy of 97.15% for this database.

For MIT-BIH NSR database no AF episodes were detected in this database. This gave us the accuracy of 100% in this database.

4 Discussion

Our goal is to develop a method which can tell whether a signal contains AF episodes or not using short length ECG signal. The heart rate variability feature of AF is exploited in this method. We have developed a novel method which uses only the scatter plot data to detect the AF. The histogram of Y axis data of scatter plot helps us to determine whether the signal contains AF episodes or not. From Figs. 3 and 4, for NSR most of the data of Y axis is concentrated in only one bin. While in case of AF (Figs. 1 and 2), data on Y axis is spread over larger region. Using the bin counts we are indirectly finding the standard deviation of Y axis data.

This method can give false results if the AF episodes are very small, i.e. of less than 6 beats. If this happens then AF episode may not get detected. To increase the accuracy further we added another method (method 2) to detect AF episodes. Fuzzy logic is useful as we can adjust the weightage to method mentioned according to the signal conditions. Since our method uses short term data to detect AF, it can easily be used for real time ECG monitoring.

From the Table 3, our algorithm provides better sensitivity and specificity than Park et al., Tateno and Glass and Dash et al. as shown. Brunn et al. has lager specificity but our algorithm has larger sensitivity. S. Ladavich, B. Ghoraani has better sensitivity but at the cost of less specificity. Thus, the proposed algorithm increases the accuracy in terms of both specificity and sensitivity.

Table 3 Comparison between various algorithm for MIT-BIH AF database

Method	Sensitivity (%)	Specificity (%)
Park et al. [13]	91.4	92.9
Tateno and Glass [9]	94.4	97.2
Dash et al. [14]	94.4	95.1
Bruun et al. [10]	96.51	99.19
Lobabi-Mirghavami et al. [15]	96.30	97.00
Ladavich and Ghoraani [6]	98.09	91.66
Proposed algorithm	96.65	97.65

5 Conclusion

This study presents a new way to detect AF using heart rate variability feature. The accuracy was of 97.23% for Physionet 2017 challenge database and 97.15% for MIT-BIH AF database. The most attractive feature of our method is detecting AF by just using the scatter plot data and histogram without calculating any statistical parameters. Our algorithm can also be used in real-time AF detection.

References

1. Wolf, P.A., et al.: Atrial fibrillation as an independent risk factor for stroke. Stroke **22**, 983–988 (1991)
2. Benjamin, E.J., et al.: Impact of atrial fibrillation on the risk of death. Am. Heart Assoc. **98**, 946–952 (1998)
3. Couceiro, R., et al.: Detection of atrial fibrillation using model-based ECG analysis. IEEE Trans. Biomed. Eng. https://doi.org/10.1109/tbme.2010.2096506
4. Du, X., et al.: A novel method for real-time atrial fibrillation detection in electrocardiograms using multiple parameters. Ann. Noninvasive Electrocardiol. https://doi.org/10.1111/anec.12111
5. Alcaraz, R., et al.: Wavelet sample entropy: a new approach to predict termination of atrial fibrillation. Comput. Cardiol. **33**, 597–600 (2006)
6. Ladavich, S., Ghoraani, B.: Rate-independent detection of atrial fibrillation by statistical modeling of atrial activity. Biomed. Signal. Process. Control **18**, 274–281 (2015)
7. Maier, C., et al.: Screening and prediction of paroxysmal atrial fibrillation by analysis of heart rate variability parameters. Comput. Cardiol. **28**, 129–132 (2001)
8. Climent, A.M., et al.: Poincare surface profiles of RR intervals: a novel noninvasive method for the evaluation of preferential AV nodal conduction during atrial fibrillation. IEEE Trans. Biomed. Eng. **56**(2), 433–442 (2009)
9. Tateno, K., Glass, L.: Automatic detection of atrial fibrillation using the coefficient of variation and density histograms of RR and ΔRR intervals. Med. Biol. Eng. Comput. **39**(6), 664–671 (2001)
10. Bruun, I.H., et al.: Automatic atrial fibrillation detection: a novel approach using discrete wavelet transform and heart rate variability. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Seogwipo, pp. 3981–3984 (2017)

11. Pan, J., Tompkins, W.J.: A real-time QRS detection algorithm. IEEE Trans. Biomed. Eng. BME **32**(3) (1985)
12. Moody, G.B., et al.: PhysioBank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation (2000). https://doi.org/10.1161/01.cir.101.23.e215
13. Park, et al.: Atrial fibrillation detection by heart rate variability in Poincare plot. Biomed. Eng. Online **8**, 38 (2009)
14. Dash, et al.: Automatic real time detection of atrial fibrillation. Ann. Biomed. Eng. **37**(9), 1701–1709 (2009)
15. Lobabi-Mirghavami, H., Abdali-Mohammadi, F., Fathi, A.: A novel grammar-based approach to atrial fibrillation arrhythmia detection for pervasive healthcare environments. JCS **2**(2) (2016)

Investigation on Daubechies Wavelet-Based Compressed Sensing Matrices for ECG Compression

Yuvraj V. Parkale and S. L. Nalbalwar

Abstract In this paper, we have investigated the different Daubechies (DB) wavelet-based compressed sensing (CS) matrices, namely db3, db4, db5, db6, db7, db8, db9, and db10 measurement matrices for ECG compression. The performance of the proposed Daubechies wavelet-based measurement matrices and state-of-the-art measurement matrices are evaluated using different performance measures such as Compression Ratio (CR), PRD, SNR, RMSE, and signal reconstruction time. The result demonstrates that the db3 and db10 measurement matrices outperform the state-of-the-art measurement matrices. Moreover, db3 and db4 measurement matrices show superior performance compared to db4, db5, db6, db7, db8, and db9 measurement matrices. Thus, this study exhibits the successful implementation of Daubechies (DB) wavelet-based sensing matrices for ECG compression.

Keywords Wavelet transform · Compressed sensing (CS) · ECG compression

1 Introduction

Compressed Sensing is a new signal compression method, in which signal is compressed at the time of acquisition. Thus, this procedure reduces the memory requirement of the system. The Donoho [1], Baraniuk [2] and Candes et al. [3] introduced the CS theory which keyed up research in different application areas such as Medical Image Processing [4], Wireless Sensor Networks [5], and ECG compression [6, 7], etc.

ECG is one of the investigative tools which provide analysis of different heart diseases. This paper presents the study of ECG signal compression using Compressed

Y. V. Parkale (✉) · S. L. Nalbalwar
Department of E & TC Engineering, Dr. Babasaheb Ambedkar Technological University (DBATU), Lonere, Raigad, India
e-mail: yuvrajparkale@gmail.com

S. L. Nalbalwar
e-mail: slnalbalwar@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_72

Sensing technique. In the literature, Polania et al. [8] applied the Random Gaussian sensing Matrix for ECG compression. Polania et al. [9] and Chae [10] implemented Bernoulli matrices for CS. Ansari-Ram et al. [11] tested a nonuniform measurement matrix. Mishra et al. [12, 13, 14] applied KLT measurement matrix and compared with the random Gaussian matrix. The result shows that the

Abbreviations	
CR—Compression ratio	SNR—Signal-to-noise ratio
PRD—Percentage root-mean-squared difference	RMSE—Root means square error

KLT measurement matrix outperforms the random Gaussian matrix. However, KLT measurement matrix is signal dependant; i.e., for different input signals, it will create different transform matrix (Φ) and thus depends upon the statistical properties of the input signal. Furthermore, the computational cost of KLT matrix is higher compared to other matrices. As a final point, in [7], the author presented an in-depth literature review on CS-based ECG compression.

In this paper, we have investigated the different Daubechies (DB) wavelet-based measurement matrices [15], namely db3, db4, db5, db6, db7, db8, db9, and db10 sensing matrices for the ECG compression. The performance of proposed DB-based measurement matrices and state-of-the-art measurement matrices are evaluated using different performance measures.

The paper is arranged as follows: Sect. 2 illustrates the Compressed Sensing (CS) theory. Section 3 describes the design steps of proposed sensing matrices. Section 4 illustrates the experimental results and discussions. Finally, the Sect. 5 present the conclusions.

2 Compressed Sensing (CS) Theory

CS is a new signal compression method, wherein the signal is compressed at the time of acquisition. Thus, this procedure reduces the memory requirement of the system. However, the successful realization of the compressed sensing is depended upon the sparsity of the signal, Restricted Isometry property (RIP), and incoherence between the sparsifying matrix (Ψ) and the sensing matrix (Φ).

Figure 1 shows the comprehensive Compressed Sensing (CS) structure. It consists of two main models, namely signal acquisition model and signals reconstruction model. In the signal acquisition model, the original signal f is sparsified using sparsifying transform matrix (Ψ) and then, this signal is compressed (y) with random sensing matrix (Φ). In the signal reconstruction model, the optimization technique such as L1-minimization [16] is used to find out the sparsest solution and then, inverse transform matrix (Ψ^*) is applied to recover the signal.

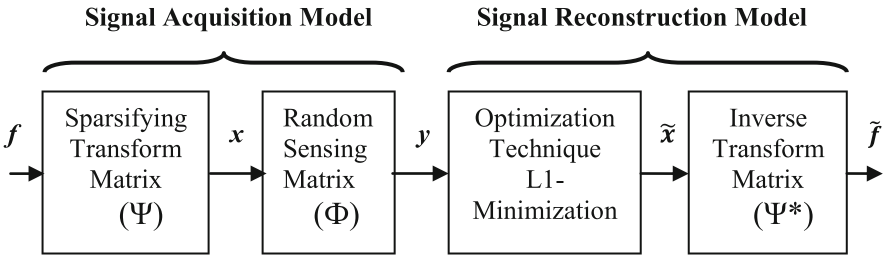


Fig. 1 Shows compression and reconstruction in compressed sensing (CS)

where, f = is the input signal, $\boldsymbol{\Psi}$ = is sparsifying matrix, $\boldsymbol{x}$ = is sparse signal under basis $\boldsymbol{\Psi}$, y = is the compressed signal under Φ, $\boldsymbol{\Phi}$ = is the sensing/measurement matrix, $\tilde{x}$ = is the sparsest sparse solution vector, and $\tilde{f}$ = is the recovered signal.

3 Constructional Steps for the Proposed Sensing Matrices

In this work, we have proposed different Daubechies (DB) wavelet-based measurement matrices. The measurement matrices are constructed using the following steps:

(1) Construct the Quadrature Mirror Filters (QMFs) with Daubechies (DB) wavelets such as db3, db4, etc.
(2) Next, construct the Identity matrix of size N × N.
(3) Then, perform 1D forward discrete wavelet transform on the Identity matrix (N × N). This will produce N × N wavelet transform matrix.
(4) Finally, select the required number of measurements (m) from N × N wavelet transform matrix.
(5) Thus, sensing matrix of size m × N is generated and used for signal compression.

4 Experimental Results and Discussions

The experimentation is conducted on ECG signal from MIT-BIH Arrhythmia Database [17]. The details of the ECG signal used are as follows: 100.dat, the number of samples selected is 1024. The DCT transform is used as sparsifying matrix. The ECG signal is reconstructed using L1-minimization technique [16]. The MATLAB 7.8.0 (R2009A) is used for the simulation along with Intel (R) Core 2 Duo CPU and 3 GB RAM PC configuration.

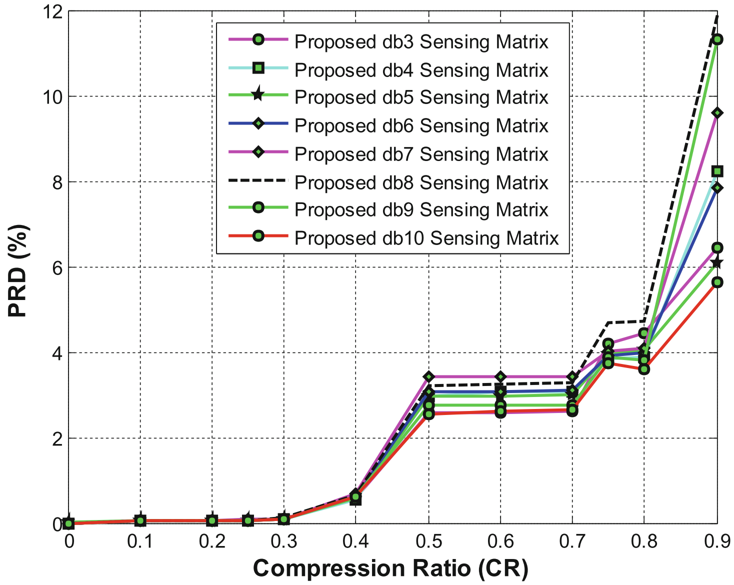


Fig. 2 Compression ratio (CR) versus PRD (%)

4.1 Performance Comparison of Different Proposed Daubechies (DB) Wavelet-Based Measurement Matrices

In this section, the performance of proposed db3, db4, db5, db6, db7, db8, db9, and db10 measurement matrices are evaluated using different performance measures such as CR, PRD, RMSE, SNR, and signal recovery time.

Figure 2 shows that db3 and db10 measurement matrices achieve lower values of PRD (%) compared to db4, db5, db6, db7, db8, and db9 sensing matrices. From Fig. 3, it is seen that db3 and db10 sensing matrices shows smaller values of RMSE compared to db4, db5, db6, db7, db8, and db9 sensing matrices. The db3 and db10 sensing matrices attain higher SNR when compared to db4, db5, db6, db7, db8, and db9 sensing matrices as shown in Fig. 4. In terms of signal recovery time (in seconds), the db3 measurement matrix requires lower signal recovery time when compared to db10 sensing matrix as shown in Fig. 5.

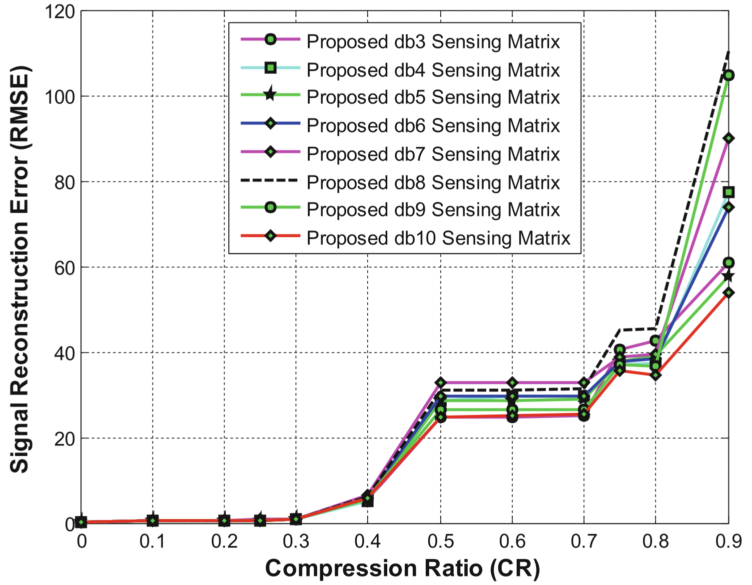


Fig. 3 Compression Ratio (CR) versus RMSE

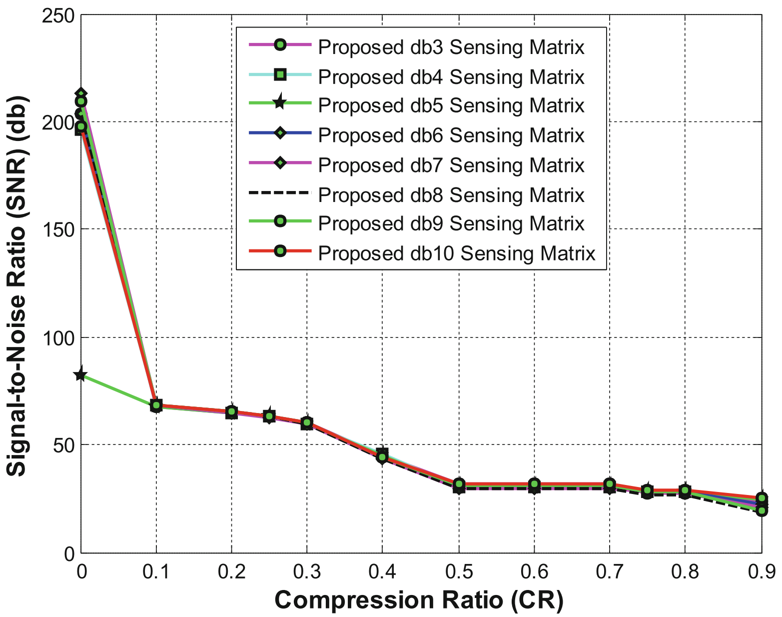


Fig. 4 Compression ratio (CR) versus SNR (db)

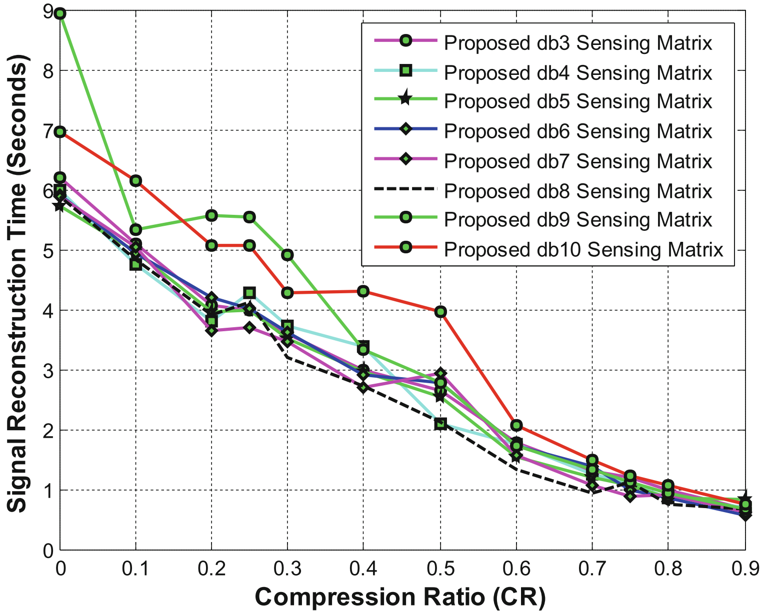


Fig. 5 CR versus signal recovery time (in seconds)

4.2 Performance Comparison of the Best Proposed Daubechies (DB) Wavelet-Based Measurement Matrices and the State-of-the-Art Measurement Matrices

This section presents the performance comparison of the best proposed Daubechies (DB) wavelet-based measurement matrices, namely db3 sensing matrix and db10 sensing matrix, and the state-of-the-art measurement matrices, namely KLT transform sensing matrix, DCT sensing matrix, and random Hadamard sensing matrix. The performance is evaluated using different performance measures such as CR, PRD, RMSE, SNR, and Signal reconstruction time.

Figure 6 shows that overall db3 and db10 measurement matrices achieve lower values of PRD (%) when compared to state-of-the-art measurement matrices such as KLT sensing matrix, DCT sensing matrix, and Circulant sensing matrix.

From Fig. 7, it is seen that overall db3 and db10 sensing matrices show smaller values of RMSE compared to state-of-the-art measurement matrices. The db3 and db10 sensing matrices attain higher SNR compared to state-of-the-art measurement matrices as shown in Fig. 8. In terms of signal recovery time (in seconds), the db3 sensing matrix requires lower signal recovery time when compared to db10 sensing matrix and state-of-the-art measurement matrices as shown in Fig. 9.

Since ECG signal posses clinically very important information; higher compression ratios are difficult to achieve in practical. Therefore, it is seen from the Fig. 10

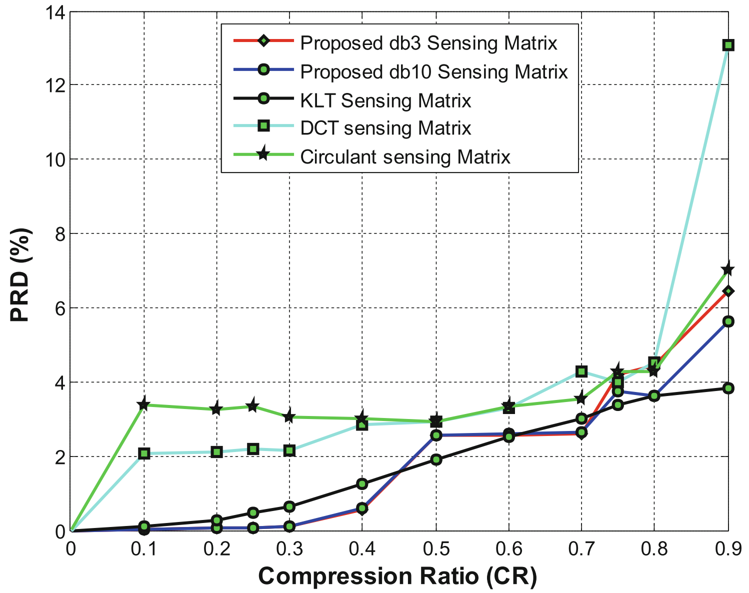


Fig. 6 CR versus PRD (%)

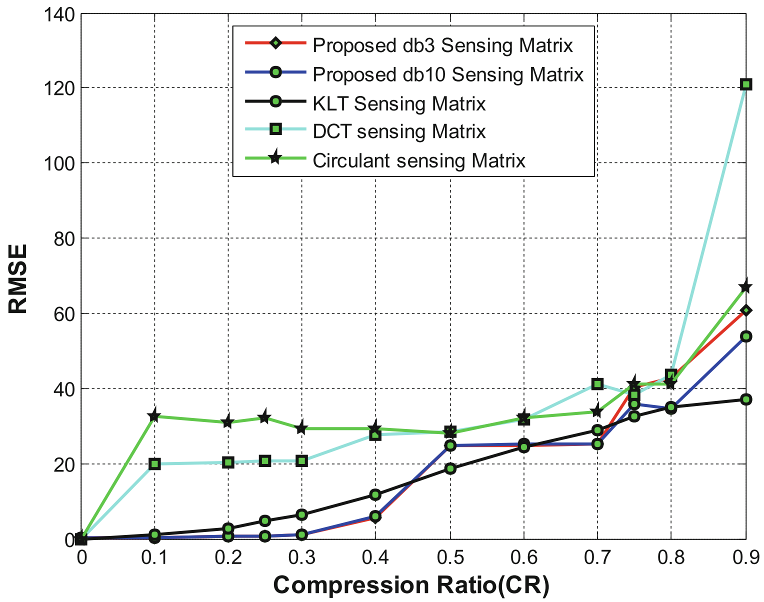


Fig. 7 Compression ratio (CR) versus RMSE

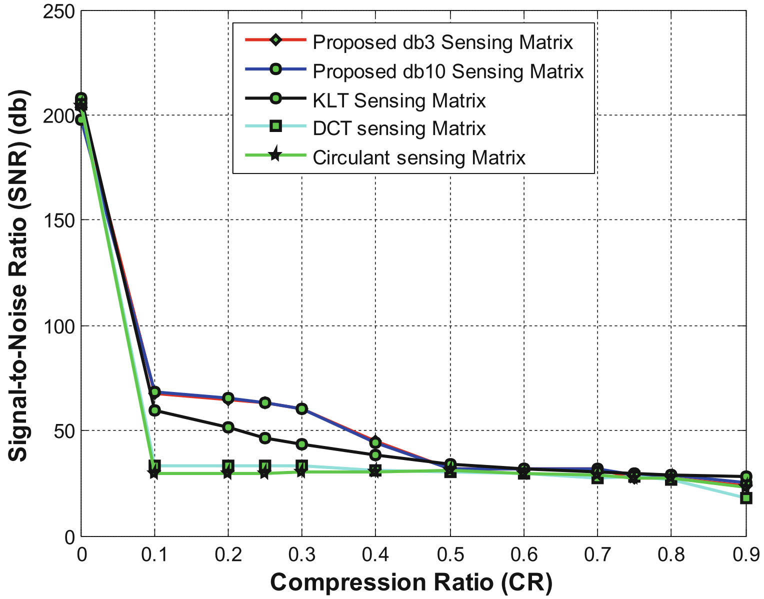


Fig. 8 Compression ratio (CR) versus SNR (db)

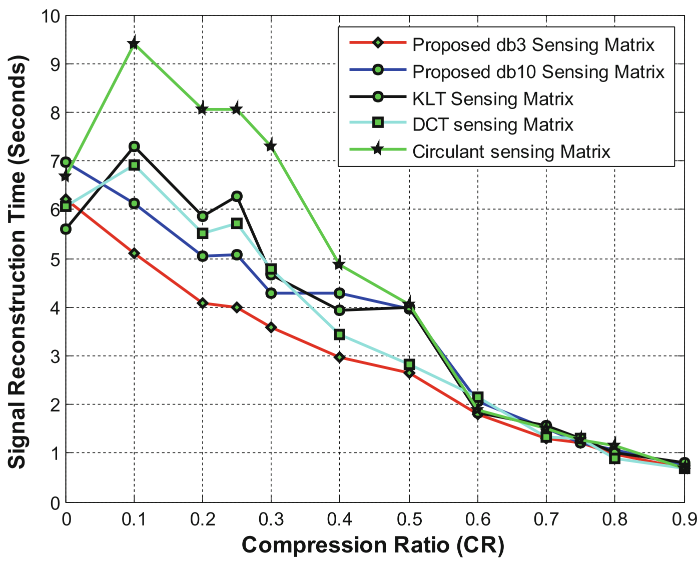


Fig. 9 CR versus signal recovery time (in seconds)

that for CR = 0.4 or 60% sampling rate (m/N = 0.6), the recovered ECG signal is very close to the original ECG signal using the proposed db3 sensing matrix.

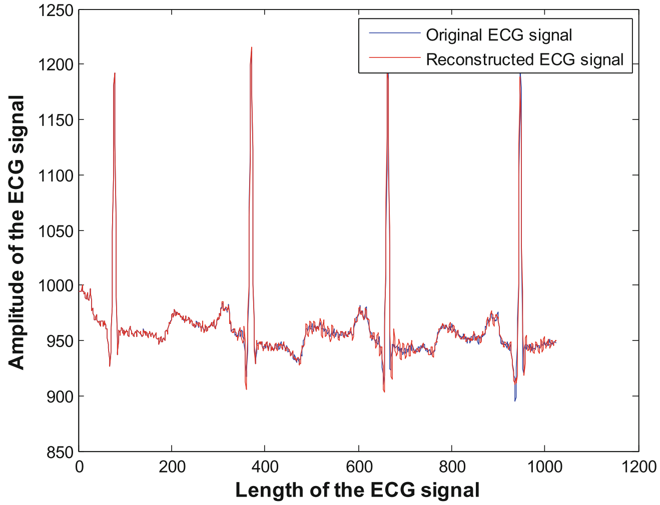


Fig. 10 Shows the original and the recovered ECG signal for 60% sampling rate using the proposed db3 sensing matrix

5 Conclusion

In this study, an attempt has been done to explore the suitability of Daubechies (DB) wavelet-based sensing matrices for ECG compression. We have proposed different Daubechies (DB) wavelet-based measurement matrices, namely db3, db4, db5, db6, db7, db8, db9, and db10 measurement matrices. The following are the major conclusions drawn from the study:

- The db3 and db10 sensing matrices achieve higher SNR and lesser values of PRD (%) along with faster reconstruction time and thus, they exhibit superior performance as compared to the db4, db5, db6, db7, db8, and db9 measurement matrices.
- Furthermore, when compared to state-of-the-art measurement matrices, the db3 and db10 sensing matrices attain higher SNR, and lower values of PRD (%) and RMSE, and thus outperform the state-of-the-art measurement matrices.

Thus, this study signifies the successful implementation of Daubechies (DB) wavelet-based measurement matrices for ECG compression. In the future prospect work, we are going to investigate the Coiflets, Symlets, and other wavelets for ECG compression.

References

1. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
2. Baraniuk, R.G.: Compressive sensing. IEEE Sig. Process. Mag. 118–121 (2007). https://doi.org/10.1109/msp.2007.4286571
3. Candes, E.J., Wakin, M.B.: An introduction to compressive sampling. IEEE Sig. Process. Mag. 21–30 (2008). https://doi.org/10.1109/msp.2007.914731
4. Lustig, M., Donoho, D.L., Santos, J.M., Pauly, J.M.: Compressed sensing MRI. IEEE Sig. Process. Mag. **25**(8), 72–82 (2008). https://doi.org/10.1109/MSP.2007.914728
5. Guan, X., Yulong, G., Chang, J., Zhang, Z.: Advances in theory of compressive sensing and applications in communication. In: Proceedings of IEEE First International Conference on Instrumentation, Measurement, Computer, Communication and Control, pp. 662–665 (2011). https://doi.org/10.1109/imccc.2011.169
6. Zhang, Z., Jung, T.-P., Makeig, S., Rao, B.D.: Compressed sensing for energy-efficient wireless telemonitoring of noninvasive fetal ECG via block sparse Bayesian learning. IEEE Trans. Biomed. Eng. **60**(2), 300–309 (2013)
7. Parkale, Y.V., Nalbalwar, S.L.: Application of compressed sensing (CS) for ECG signal compression: a review. Springer-Advances in Intelligent Systems and Computing (Springer-AISC), vol. 469, pp. 53–65. https://doi.org/10.1007/978-981-10-1678-3_5
8. Polania, L.F., Carrillo, R.E., Blanco-Velasco, M., Barner, K.E.: Compressed sensing based method for ECG compression. In: Proceedings of ICASSP, pp. 761–764 (2011)
9. Polania, L.F., Carrillo, R.E., Blanco-Velasco, M., Barner, K.E.: Exploiting prior knowledge in compressed sensing wireless ECG systems. IEEE J. Biomed. Health Inform. (2014)
10. Chae, D.H., Alem, Y.F., Durrani, S., Kennedy, R.A.: Performance study of compressive sampling for ECG signal compression in noisy and varying sparsity acquisition. In: Proceedings of ICASSP, pp. 1306–1309 (2013)
11. Ansari-Ram, F., Hosseini-Khayat, S.: ECG signal compression using compressed sensing with nonuniform binary matrices. In: 16th CSI International Symposium on, Artificial Intelligence and Signal Processing (AISP), pp. 305–309 (2012)
12. Mishra, A., Thakkar, F., Modi, C., Kher, R.: ECG signal compression using compressive sensing and wavelet transform. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3404–3407 (2012)
13. Mishra, A., Thakkar, F.N., Modi, C., Kher, R.: Selecting the most favorable wavelet for compressing ECG signals using compressive sensing approach. International Conference on Communication Systems and Network Technologies (CSNT), pp. 128–132 (2012)
14. Mishra, A., Thakkar, F., Modi, C., Kher, R.: Comparative analysis of wavelet basis functions for ECG signal compression through compressive sensing. Int. J. Comput. Sci. Telecommun. **3**, 9 (2012)
15. Parkale, Y.V., Nalbalwar, S.L.: Application of 1-D discrete wavelet transform based compressed sensing matrices for speech compression. J. SpringerPlus **5**(1), 1–60 (2016). https://doi.org/10.1186/s40064-016-3740-x
16. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM Rev. **43**(1), 129–159 (2001)
17. MIT-BIH Arrhythmia Database. www.physionet.org. https://doi.org/10.13026/c2f305

Statistical Characterization of an Underwater Channel in a Tropical Shallow Freshwater Lake System

Jyoti A. Sadalage, Arnab Das and Yashwant Joshi

Abstract Underwater acoustics has made significant strides over the last century, which finds applications over a wide range from basic bathymetry study to high-end research extensions. The acoustic propagation in underwater is typically governed by physical properties of the underwater channel, such as temperature, pressure, and salinity. The seasonal fluctuations in the physical properties of the tropical region manifest as thermal stratification. The random thermal stratification has a significant impact on the Sound Speed Profile (SSP), thereby distorting the received echoes from the surface and the bottom. The site-specific behavior in the tropical region makes it an interesting research problem to investigate the correlation of the surface parameters like temperature with the surface and bottom reflection due to variations in the SSP. In this work, we attempt to present underwater channel characteristics of the tropical freshwater lake system at Khadakwasla (18.43° N, 73.76° E), located in the municipal limits of Pune city in India. The temperature gradient along the water column is computed using the one-dimensional Freshwater Lake Model (FLake) to derive the SSP using Medwin relation. The statistical analysis of the sound speed fluctuations resulted due to seasonal variation in the water temperature is presented using the Kolmogorov–Smirnov (KS) Goodness-of-Fit test is used to find a close Probability Density Function (pdf) match for the surface and the bottom path impulse response. The results indicate a good match of the surface and bottom path impulse response with Weibull distribution with a high confidence level. Such characterization can facilitate the design of adaptive algorithms to minimize the underwater channel impact based on a precise estimate of the channel impulse response.

Keywords Tropical shallow freshwater · Underwater channel impulse response Statistical characteristics

J. A. Sadalage (✉) · Y. Joshi
Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, India
e-mail: jyoti.rangole@vpkbiet.org; jyotika2k1@gmail.com

A. Das
Maritime Research Center, Pune, India

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_73

1 Introduction

The freshwater availability in the developing world is increasingly getting limited [1] and on the other hand, socioeconomic pressure due to increasing population is only on the rise. The freshwater systems like lakes act as dynamic response systems that integrate the environment, climate, and tectonic forces into a continuous, high-resolution archive of local and regional changes [2]. The tropical regions witness seasonal variation in the heat gain/loss and water depths along with the amount of throughput. Such variations have a significant impact on the thermal stratification [3] that manifest in the limnology [4]. The acoustic propagation in the tropical region witnesses twofold challenge while ensuring optimal sonar performance for an underwater application. The first is the perpetual shallow water behavior due to time-varying multipath [5] interaction with the two boundaries. The second challenge pertains to the random surface fluctuations due to diurnal and seasonal temperature and other variation that impact the thermal stratification, thereby modifying the Sound Speed Profile (SSP). The quantification of the surface parameters like the temperature can be used to assess the variation in the SSP to further model the underwater channel impact on the echo signal at the receiver. Modeling and simulation efforts to characterize the underwater channel impulse response have been used in many applications to reduce the medium impact on the sonar performance.

In this work, we attempt to present the seasonal variations in the surface temperature for a tropical freshwater system at Khadakwasla Lake [6], in India. The thermal gradient is computed, based on the surface temperature to derive the SSP that facilitates the underwater channel characterization. The channel characterization includes computation of the surface and bottom impulse response. The seasonal variations are statistically analyzed and a close match to known Probability Density Functions (pdfs) [7, 8] derived. The efforts will facilitate the understanding of the medium fluctuation on the acoustic propagation and possible design of algorithms to mitigate the impact on the received echo at the sonar receiver. Effective adaptive algorithms can substantially optimize sonar performance in the tropical regions.

The temperature gradient along the water column for the Khadakwasla Lake is obtained using the one-dimensional Freshwater Lake Model (FLake) [9–11]. The detailed seasonal variations have been recorded using the FLake model in the site for the entire year using average temperature values. The temperature gradient information is fed to the Medwin [12] formula to compute the SSP which is fed to the acoustic channel simulator [13]. The underwater channel simulator model is used to derive the surface and bottom impulse response. This paper presents the use of a large-scale model of the underwater acoustic simulator [13] that allows uncertainty in the channel geometry. The medium and the physical parameters of the Khadakwasla Lake are fed to the channel model to obtain the channel parameters. The physical parameters include the channel geometry such as water depth, transmitter and receiver depth, bottom type, etc., whereas the environmental parameter includes the water SSP. The random fluctuations of the surface and bottom impulse response have been captured in the pdf. The Goodness-of-fit test like the Kolmogorov–Smirnov (KS) test has been then used to find the close match to known pdfs.

The simulation study on real surface temperature data presents encouraging results that match the known distributions for the surface, bottom reflected paths. In the absence of a known benchmark and the wide variation in the statistical properties of the underwater channel, here we use a battery of pdfs for comparison. The surface and bottom reflected path impulse response matches with the three-parameter Weibull distribution with a confidence level of 98%. The results presented in this paper are for the Khadakwasla Lake, which supplies fresh water to the city of Pune for domestic purpose and has significant socioeconomic relevance for tropical regions in the developing world. The site has been so chosen to analyze the acoustic complexity of a tropical shallow freshwater system.

2 Data and Measurement

2.1 Location Details

Khadakwasla Lake is located at Pune city, in the state of Maharashtra (India) which spans a length of 17 km and has a width of 1 km. The total catchment area of the Lake is about 501 square km and its depth varies between 12 and 36 m [6]. Khadakwasla Lake is one of the main reservoirs meeting the freshwater requirements of the Pune city.

2.2 Freshwater Lake Model

FLake [9–11] is the freshwater Lake model capable of predicting vertical temperature structure and mixing condition of the Lake of various depths on time scales from few hours to few years. FLake model is based on a two-layer parametric representation of the evolving temperature profile. The structure of the stratified layer between the upper mixed layer, the basin bottom, and the Lake thermocline, is based on the concept of self-similarity of the temperature-depth curve. Month-wise temperature variation derived from the FLAKE model at Khadakwasla Lake for the year 2005–2006 has been presented in Fig. 1. The absence of real temperature data at Khadakwasla Lake compelled us to use the open source version of the Flake-Global model [9] for calculation of approximate water temperature. The model gives an estimate of the surface and bottom water temperatures, mean temperature of the water column, surface mixed layer depth using the geographical coordinates, mean depth to the bottom, and water transparency of the lake. The model adopts the metrological parameters such as air temperature, wind speed, relative humidity and the solar radiation from Global Data Assimilation System. The model results indicate signif-

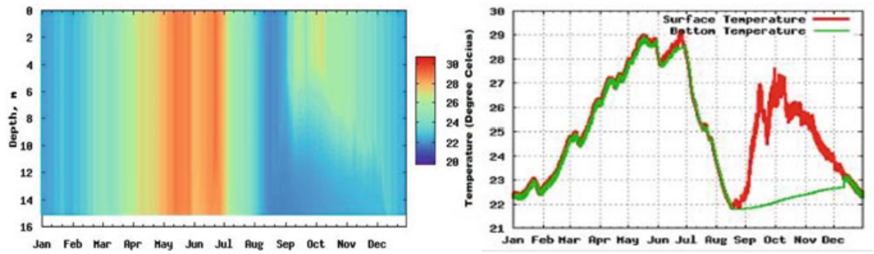


Fig. 1 Month-wise water temperature variation at Khadakwasla lake

icant variation in the temperature gradient along the water column for the month of October and November, attributable to the tropical littoral behavior.

2.3 Sound Speed Profile (SSP)

The underwater medium from the acoustic perspective is typically defined based on the SSP that impacts the propagation of the acoustic signal. Tropical waters present diurnal and seasonal temperature variation and are highly site-specific. Some of the empirical formulas for SSP that conform to the tropical shallow freshwater conditions include Wilson [14], Medwin [12], Chen and Millero [15], Coppens [16], Leroy [17]. This paper uses Medwin's relation as shown in Eq. (1)

$$C = 1449.2 + 4.6T - 0.055T^2 + 0.00029T^3 + (1.34 - 0.010T)(S - 35) + 0.016D \tag{1}$$

where T—Temperature in °C, S—Salinity in ppt, D—Depth in m.

FLake model is found to be effective in providing the temperature gradient along the water column. The authors have field validated the FLake model in the same site and the same is been reported in a separate publication. The salinity change along the water column has not been taken into consideration. The average salinity considered here is 0.2 ppt [18]. Month-wise variation in the sound speed along the water column of Khadakwasla Lake of the year 2005–2006 is presented in Fig. 2. We could see a corresponding variation in SSP approximately from 1488 to 1506 m/s across the year as shown in Fig. 2. This would result in significant impact on the signal propagation.

2.4 Channel Impulse Response

The multipath phenomenon in underwater acoustic communication ensures different path modifications and delays to the propagating source signal. Every path will have

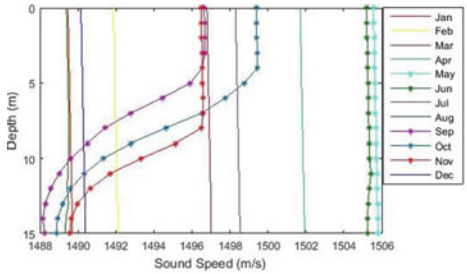


Fig. 2 Month-wise variation of sound speed along the water column

its own frequency response that is characterized by its channel geometry, reflection, and refraction properties of the medium. The multipath channel is typically assumed to have a finite number of paths. Let l_p and τp denote path length and path delay of the pth path respectively. The transfer function of the pth path is given by Eq. (2) [19]

$$Hp(f) = \frac{\Gamma p}{\sqrt{A(lp, f)}} \tag{2}$$

Γp indicates the cumulative reflection coefficient along the pth path and $A(lp, f)$ indicates the propagation loss as given in Eq. (3) [19]

$$A(lp, f) = \left(\frac{l}{lr}\right)^k a(f)^{(l-lr)} \tag{3}$$

$a(f)$ indicates the absorption coefficient given by empirical formula [20], l_r indicates the reference path length. Each path of the acoustic channel is considered to be a low pass filter which contributes to the overall impulse response of the channel. Equation (4) shows the overall channel impulse response [19].

$$h(t) = \sum_p hp(t - \tau p) \tag{4}$$

where $hp(t)$ is the Inverse Fourier Transform of $Hp(f)$.

3 Related Work and Simulation Efforts

Unlike radio channels where a number of models for the probability distribution are well accepted and even standardized, there is limited consensus on statistical characterization of underwater acoustic channels [19]. Significant literature does exist on underwater communication work, though it may not be directly related to the proposed work in terms of application, however, similarities can be drawn on the propagation analysis and reasonable inferences concluded. Some authors find Ricean fading or Rayleigh fading good approximation for their data [20, 21]. Hovannes et al. [22] presented a statistical characterization of shallow and very shallow water communication. The KL divergence Goodness-of-fit test results presented by Hovannes, K. et al. confirm that very shallow water impulse response does not necessarily follow Rayleigh distribution rather exhibits close to Weibull or Rice distribution and shallow water environment matches close to beta distribution. Yang [23] confirms to the K distribution fading for the underwater channel. Jian et al. [24] presented a statistical characterization of the underwater acoustic channel at Narragansett Bay. The KS test employed by Jian Zhang concluded that the magnitude of channel impulse response PDF matches the compound K distribution. The literature presents the variety of proposed models; this would be due to the site-specific and experiment-specific properties. Taking this fact into account, the proposed study attempts to analyze the impulse response characteristics of a typical tropical shallow freshwater lake site at Khadakwasla.

Use of large-scale underwater acoustic model proposed by Qarabaqi et al. [13] allowed us to take under consideration the change in environmental condition such as sound speed and channel geometry including the transmitter height, receiver height, distance between transmitter and receiver, etc., over a period of time. The simulation is carried out for a channel geometry, with the water column of an average depth of 25 m, transmitter at depth of 15 m, receiver at depth of 15 m, and cylindrical spreading factor ($k = 1$). The underwater acoustic channel model is simulated number of times to allow the uncertainty in channel geometry. The transmitter and receiver depth are allowed to vary by ± 0.25 m. The distance between transmitter and receiver is set at 100 m and allowed to vary by ± 0.25 m. The time evolution in the magnitude of the multipath impulse response for minimum and maximum sound speed of 1488 and 1506 m/s, could be visualized from Fig. 3a and Fig. 3b, respectively. The X-axis represents the time of arrival (in ms) of each path, while Y-axis represents the instantaneous time. The magnitude of the surface and the bottom path response with variable water sound speed are further analyzed for verification of the pdf of the surface and bottom path impulse response. The minimum operating frequency and the bandwidth selected is 10 kHz. The channel impulse response is simulated for duration of 180 s. Figure 3a, b shows the magnified version of the multipath impulse response in order to clearly locate the surface and bottom path impulse response. Po represents the direct path, Ps represents the surface reflected path, Pb represents the bottom reflected the path of the impulse response, and s1b1 indicates the signal with one surface and one bottom bounce.

4 Results and Discussion

The objective of the study is to present acoustic characteristics of the tropical shallow freshwater channel in terms of its SSP and multipath impulse response. The proposed work presents the statistical characterization of the impulse response of Khadakwasla

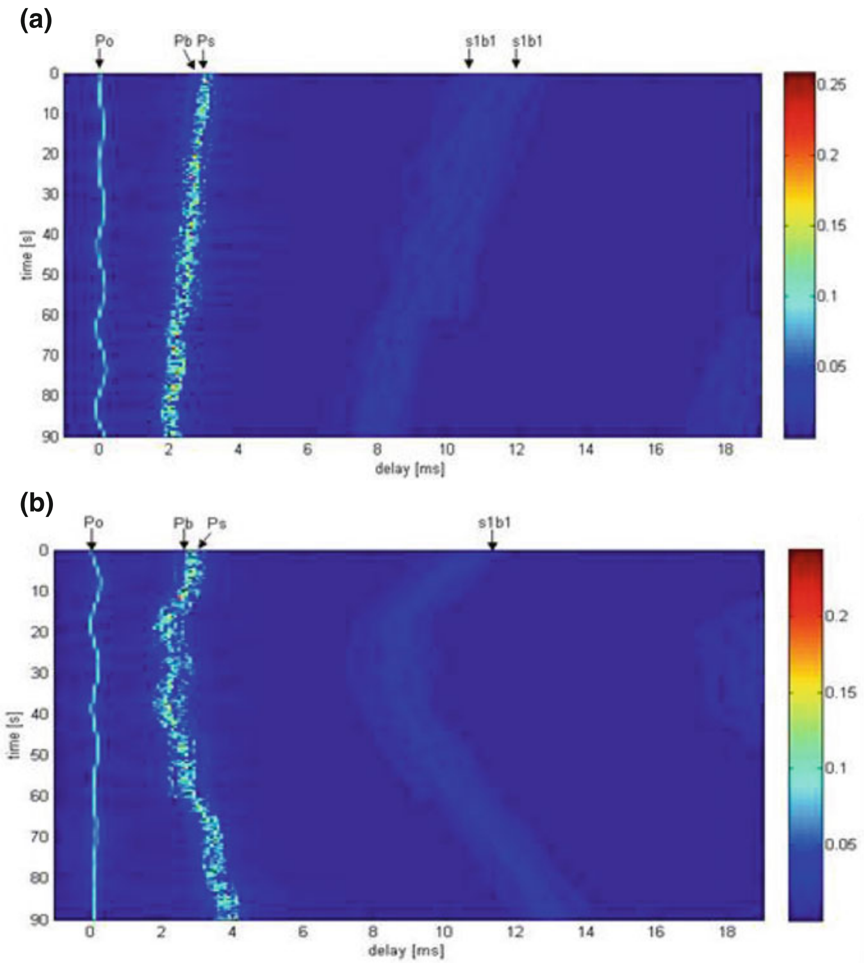


Fig. 3 Time evolution of magnitude baseband impulse response for water sound speed of **a** 1488 m/s **b** 1506 m/s

Lake in terms of its pdf distribution. The variation in channel geometry and water sound speed is also incorporated for evaluating the best fit for the channel impulse response. The acoustic channel model is simulated multiple times for obtaining the representative surface and bottom path impulse response over a specific range of variations in the channel geometry. The surface and bottom path impulse response have been analyzed in chunks of 25 s. The pdf fitting for the surface and bottom path impulse response is presented in Fig. 4a and Fig. 4b, respectively, that does reflect consonance with the standard trends in the literature on pdf characteristics of the underwater channels. The KS Goodness-of-fit test is carried out for five different distributions, namely Normal, Lognormal, Rayleigh, Rice, and Weibull (3P) for both

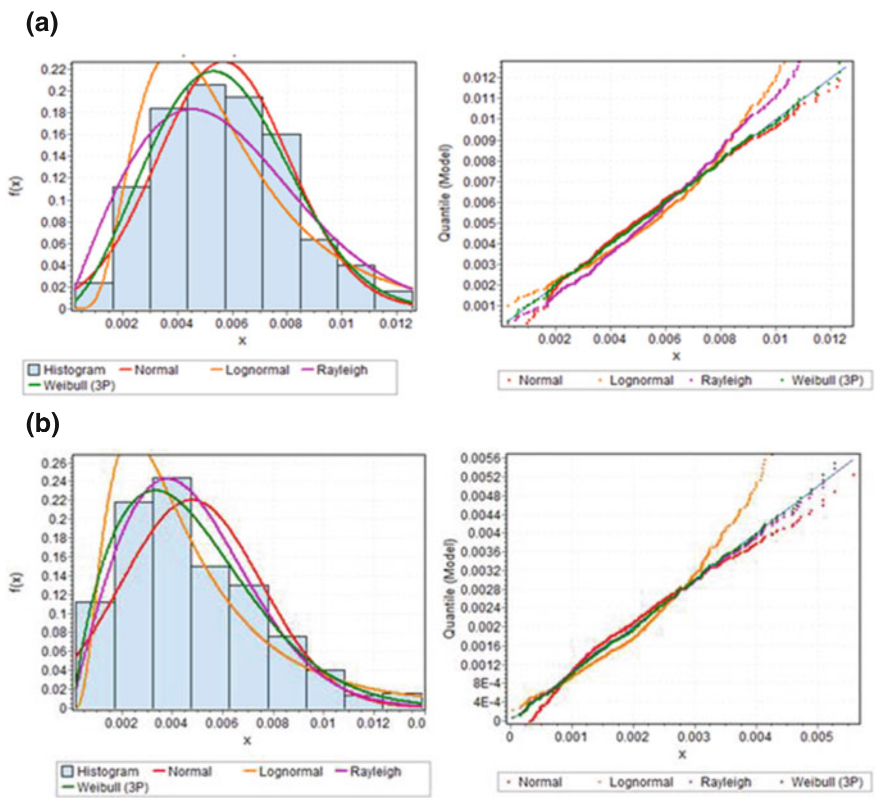


Fig. 4 Curve fitting of standard pdf to the impulse responses. **a** Surface reflected path. **b** Bottom reflected path

surface and bottom reflected paths. This test is based on the empirical cumulative distribution function and has been used to decide, if the sample comes from the hypothesized continuous distribution. The null hypothesis is defined as H0: *The impulse response follows specified distribution* while the alternative hypothesis is defined as H_A: *The impulse response does not follow specified distribution.* The test statistics (D value) is based on the largest vertical difference between the theoretical and empirical cumulative distribution [8]. The null hypothesis is tested at significance value of 0.02 and 0.05. Table 1 and Table 2 present the summary of the test statistics for the surface and bottom path response, respectively. The result shows that the Weibull (3P) distribution fits both surface and bottom path impulse response with the confidence level of 98%. The Rayleigh and Rice distribution outperforms for a few cases with confidence level of 95%, whereas the Normal and Lognormal distribution shows poor fitting for both the paths.

The Quantile-Quantile (Q–Q) plot shows the graph of the impulse amplitude plotted against the theoretical (fitted) distribution quantiles. The Q–Q plot clearly

Table 1 Summary of KS Goodness-of-fit test statistics for surface reflected path

Name of distribution	Surface reflected path 1	Surface reflected path 2	Surface reflected path 3	Surface reflected path 4	Surface reflected path 5
Weibull (3P)	0.0223	0.0271	0.0281	0.031	0.0351
Rayleigh	0.0228	0.0360	0.0286	0.0373	0.0395
Rice	0.0389	0.0372	0.0365	0.0494	0.0481
Normal	0.0498	0.0441	0.0560	0.0491	0.053
Lognormal	0.0898	0.0841	0.0953	0.0923	0.0923

Table 2 Summary of KS Goodness-of-fit test statistics for bottom reflected path

Name of distribution	Bottom reflected path 1	Bottom reflected path 2	Bottom reflected path 3	Bottom reflected path 4	Bottom reflected path 5
Weibull (3P)	0.0261	0.0208	0.0315	0.0271	0.0266
Rayleigh	0.0277	0.0217	0.0373	0.0360	0.0280
Rice	0.0316	0.0395	0.0494	0.0372	0.0308
Normal	0.0633	0.056	0.0491	0.0441	0.0009
Lognormal	0.0791	0.0793	0.0923	0.0841	0.0627

shows that the Weibull (3P) shows good fit to both the surface and the bottom path impulse response as compared to the other distributions. The normal and lognormal distribution shows the skewness in the fitted data and hence, it would not be a good choice to go with.

5 Conclusion

The tropical shallow water behavior has typically been considered to be highly complex and relatively less studied. The freshwater system located in the tropical shallow water region exhibits typical behavior with high random variations that result in the suboptimal performance of the SONAR system deployed The annual variations in the surface temperature translate to fluctuations in the medium properties manifested in the form of sound speed in the medium. The results of the simulation effort at a realistic experimental site indicate significant fluctuations in the seasonal parameters. Use of FLake model has given an insight into the temperature gradient along the water column which could result in the change in the sound speed along the water column. The large-scale acoustic channel model allowed imparting uncertainty in the channel geometry over a period of time. The KS Goodness-of-fit test shows that the surface and bottom path impulse response of shallow freshwater channels present a close match to the three parameters Weibull distribution with the confidence level of 98%. The simulation efforts give us an enhanced understanding of the entire process of medium fluctuation at tropical shallow water along with characterization of channel impulse response. The impact of water column gradient on the surface and bottom reflection is seen for the month of September, October, and November.

References

1. United Nations. Water for life 2005–2015. http://www.un.org/waterforlifedecade/scarcity.html. Accessed 12 Sept 2017
2. Gierlowski, E.H., Kelts, K.: Lake basins through space and time. AAPG Studies in Geology. American Association of Petroleum Geologists, vol. 46, p. 648, Tulsa, OK (2000)
3. Bertram, B., Marti, S.: Stratification of lakes. Rev. Geophys. **46**, RG2005 (2008)
4. William, L.: Basics for protection and management of tropical lakes. Lakes Reserv. Res. Manag. **5**, 35–48 (2000)
5. Etter, P.C.: Underwater Acoustic Modeling and Simulation, 3rd edn. Spon Press
6. National Defence Academy. www.nda.nic.in. Accessed 23 May 2017
7. Walck, C.: Handbook on statistical distributions for experimentalists. Particle Physics Group, Fysikum, University of Stockholm
8. http://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/lecture14.pdf. Accessed 03 Sept 2016
9. Kirillin, G., Hochschild, J., Mironov, D., Terzhevik, A., Golosov, S., Nützmann, G.: FLake-global: online lake model with worldwide coverage. Environ. Model Softw. **26**(5), 683–684 (2011)
10. Thiery, W., Martynov, A., Darchambeau, F., Descy, J., Plisnier, P., Sushama, L., Lipzig, V.: Understanding performance of Flake model over two African great lakes. Geosci. Model Dev. **7**, 317–337 (2014)
11. Georgiy, K., Lijuan, W., Tom, S.: Seasonal thermal regime and climatic trends in lakes of Tibetan highlands. Hydrol. Earth Syst. Sci. (2016)
12. Medwin, H.: Speed of sound in water: a simple equation for realistic parameters. J. Acoust. Soc. Am. **58**(6), 1318–1319 (1975)
13. Qarabaqi, P., Stojanovic, M.: Statistical characterization and computationally efficient modeling of a class of underwater acoustic communication channels. IEEE J. Ocean Eng. **38**(4) (2013)
14. Wilson, W.D.: Equation for the speed of sound in sea water. J. Acoust. Soc. Am. **32**(10), 1357 (1960)
15. Chen, C.-T., Millero, F.J.: Speed of sound in seawater at high pressures. J. Acoust. Soc. Am. **62**(5), 1129–1135 (1977)
16. Coppens, A.B.: Simple equations for the speed of sound in Neptunian waters. J. Acoust. Soc. Am. **69**(3), 862–863 (1981)
17. Leroy, C., Stephen, P., Goldsmith, M.: A new equation for the accurate calculation of sound speed in all oceans. J. Acoust. Soc. Am. **124**(5), 2774–2782 (2008)
18. Twomey, L., Piehler, M., Paerl, H.: Priority parameters for monitoring of fresh water and marine system and their measurement. Environmental monitoring. Encyclopedia of Life Support Systems
19. Stojanovic, M., Preisig, J.: Underwater acoustic communication channels: propagation models and statistical characterization. IEEE Commun. Mag. (2009)
20. Chitre, M.: A high-frequency warm shallow water acoustic communications channel model and measurements. J. Acoust. Soc. Am. **122**(5), 2580–2586 (2007)
21. Radosevic, A., Proakis, J., Stojanovic, M.: Statistical characterization and capacity of shallow water acoustic channels. In: IEEE Oceans Europe Conference (2009)
22. Hovannes, K., Melodia, T.: Modeling underwater acoustic channels in short-range shallow water environments. In: Proceedings of WUWNet'14, Rome, Italy, 12–14 Nov 2014

23. Yang, W., Yang, T.C.: High-frequency channel characterization for m-ary frequency-shift-keying underwater acoustic communications. J. Acoust. Soc. Am. **120**(5), 2615–2626 (2006)
24. Jian, Z., Jesse, C., Yahong, Z.: Statistical channel modeling of wireless shallow water acoustic communications from experiment data. In: Military Communications Conference (2010)

Nonuniform Frequency Sampling Approach to FIR Filter Design

Mahesh Ladekar, Yashwant Joshi and Ramchandra Manthalkar

Abstract This paper investigates the new approach to FIR filter design based on nonuniform frequency sampling. This method generates the nonuniform samples in passband and stopband separately using Gaussian function. For the generated nonuniform sample, the desired frequency response values are generated using ideal filter characteristics. Then, taking its nonuniform IDFT gives the required filter coefficients. The proposed method is compared with existing methods like uniform frequency sampling and optimal filter design method and results show that the investigated approach has a better advantage over uniform frequency sampling and Parks–McClellan method with regard to the frequency response of designed filter.

Keywords Finite impulse response (FIR) · Infinite impulse response (IIR)
Nonuniform discrete fourier transform (NDFT)

1 Introduction

The digital filter is a most widely used digital signal processing block. There are two types of digital filters as finite-duration impulse response (FIR) and infinite duration impulse response (IIR). FIR filters are important in many applications over IIR due to several advantages. These advantages include the exact linear phase, absolute stability and ease for implementation. This paper presents the application of NonUniform Discrete Fourier Transform (NDFT) in FIR filter design.

Most widely used techniques for FIR filter design are window-based approach, frequency sampling and Parks–McClellan designs [1]. The filter designed by window-based method and frequency sampling method are computationally simple and easy to design but suffer from disadvantages of Gibbs phenomenon and ripples between sample values of the realized filter.

M. Ladekar (✉) · Y. Joshi · R. Manthalkar
Department of Electronics & Telecommunication, Shri Guru Gobind Singhji Institute of Engineering & Technology, Nanded 431606, Maharashtra, India
e-mail: mladekar@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_74

In [2], the basic frequency sampling method for FIR filter design, the desired frequency response is obtained by sampling ideal frequency response at N equally spaced frequencies, where N is the length of filter. The values at these uniform frequency samples are treated as Discrete Fourier Transform (DFT) coefficients. By taking inverse DFT, the filter coefficients are computed. Thus, the frequency response of the resulting filter is the interpolated response between the frequency samples. This method satisfies the frequency response values at specified frequencies exactly but has large approximation error at intermediate frequencies. The interpolated response has large ripple at the edge of passband due to sudden change in desired response. So, to improve the filter characteristic, transition band is added in the ideal frequency response.

The frequency sampling approach to FIR filter design is developed further by Rabinar et al. [3]. In this, a number of samples of transition band are varied in amplitude such that the maximum deviation from the desired response is minimum. Linear Programming is used to optimize the values of these variable samples. Hence, the method is computationally inefficient.

In [3], the technique for nonuniform frequency sampling (NUFS) was included. In which, the design of low-pass filter using two different sets of nonuniform data are used. The first set consists of uniform samples with an extra sample in transition band. And, the second set is of two extra samples in transition band. The desired frequency characteristic is one in the passband and zero in the stopband. For extra sample placed, the optimum value was taken, so that the peak ripple is minimum in stopband. But in both cases, peak ripples is observed and it does not give major advantage. After this, Angelides et al. [4] have investigated further on nonuniform frequency sampling. Newton-type polynomial with complex conjugate coefficients is used for interpolating the desired frequency response of a real coefficient FIR filter.

The use of transition functions as pth order spline function in the ideal frequency response is proposed by Burrrus et al. [5]. The transition function is added in the frequency characteristics and effect of that acts as weighting function in time domain to reduce the Gibbs' phenomenon.

In [6, 7], the nonuniform frequency sampling approach to FIR filter design is given an application of NDFT. In this method, the design procedure is divided into two steps, the first step is generation of desired frequency response and the second is selection of frequency-sample locations. In the first step, for given filter specification, desired frequency response is constructed using analytic function in passband and stopband separately. The Chebyshev polynomial is used as analytic function to get equiripple frequency response. In second step, the frequency samples are selected as the extrema of the desired equiripple response. Using NDFT matrix constructed from selected frequency location and desired frequency response in the first step forms system of linear equations. By solving this system of linear equations, the filter coefficients are obtained. The obtained filter is nearly equal to optimal equiripple filter and design time is reduced as compared to Parks–McClellan algorithm.

Thus, in designing digital filter using nonuniform frequency sampling approach, two major issues are involved, i.e., generating desired frequency response and choice of frequency location. In the proposed method for designing low-pass FIR filter, the desired frequency response was considered one in passband and zero in the stopband. And, the investigation was carried out for generating nonuniform sample points in frequency domain. In the literature, [8–12] suggested the approaches for generating nonuniform sample points in case of sampling of time domain signal. And, the formulation for reconstruction from nonuniform samples was also discussed. Four different methods for non-uniform sampling were suggested,

1. Migration of a Finite Number of Uniform Sample Points
2. Jitter Sampling
3. Periodic Nonuniform Sampling
4. Chebyshev Nonuniform Sampling.

First approach is one of the simplest examples of nonuniform sampling. In jitter sampling, samples are selected around uniform samples either deterministically or randomly with a given probability distribution. In periodic non-uniform sampling, nonuniform samples are periodic in some sense and generated by combination of different kind of uniform samples. In Chebyshev, nonuniform sampling points are generated by the roots of Nth degree Chebyshev polynomial. These approaches were investigated to choose the location of nonuniform frequencies for designing but the results were not encouraging.

2 Nonuniform Discrete Fourier Transform

In DSP, the frequency-domain representation of discrete sequence is Discrete Fourier Transform which is sampled version of discrete-time Fourier transform. In most practical signals, the energy is nonuniformly distributed in frequency domain. So, the concept of NDFT is introduced by [6, 7]. The NDFT is defined for a finite length sequence as samples of its Z-transform evaluated at arbitrarily chosen points in the z-plane.

Thus, in [7], the nonuniform discrete Fourier transform of a sequence $x[n]$ of length N is defined as

$$X(z_k) = \sum_{n=0}^{N-1} x[n] z_k^{-n},\ k = 0, 1, \ldots, N-1 \tag{1}$$

where $z_0, z_1, \ldots, z_{N-1}$ are distinct points located arbitrarily in the z-plane on the unit circle. Here, computing inverse NDFT can be viewed as polynomial interpolation problem. So, the existing methods for polynomial interpolation like Langrange Interpolation and Newton Interpolation can be used to solve inverse NDFT equation.

3 Gaussian Window Function

In the proposed method, we use Gaussian window function to generate nonuniform frequency samples for the design. The Gaussian window function is defined in [13] as

$$w(n) = \exp\left[\frac{1}{2}\left[\alpha\frac{n}{N/2}\right]^2\right] \tag{2}$$

where N is the length of window and α is the reciprocal of standard deviation, a measure of the width of its Fourier transform. Increased α will decrease the width of the window and reduce the discontinuity at the boundaries.

4 Proposed Method

A new method is proposed for FIR filter design based on nonuniform frequency sampling. The proposed method is better than the uniform frequency sampling approach and nearly approximates the well-known Parks–McClellan Algorithm with regard to frequency response of designed filter.

In the proposed method, the desired ideal frequency response is sampled at N non-equispaced points in the frequency domain. An N point inverse NDFT of these samples gives the filter coefficients. Two problems are envisaged before designing the filter. One is the generation of desired frequency response and generation of nonuniform frequency samples. The aim of the work is to obtain the filter which has interpolated frequency response satisfying given specifications.

Consider the design of a linear phase low pass filter of even length and symmetric impulse response. The frequency response of this filter is

$$H\left(e^{j\omega}\right) = B(\omega)e^{-j\omega(N-1)/2} \tag{3}$$

where $B(\omega)$ is an even and periodic function of ω, is given by

$$B(\omega) = \sum_{n=0}^{N/2} b[n]\cos\left[\omega\left(n - \frac{1}{2}\right)\right] \tag{4}$$

and

$$b[n] = 2h\left[\frac{N}{2} - 1\right],\ n = 1, 2, \ldots, \frac{N}{2} \tag{5}$$

The procedure of designing low-pass filter is carried out by first solving two problems mentioned. (1) Generation of desired frequency response: The desired frequency response is obtained by sampling of ideal low pass filter frequency response. (2) Generation of nonuniform frequency samples: The nonuniform frequency samples are generated separately in passband and stopband by a Gaussian window function.

Let the filter has passband edge at ω_p and stopband edge at ω_s and passband ripples as δ_p and stopband ripples as δ_s. In the passband, the nonuniform samples are selected by an equation [6] as

$$\omega_k = \omega_p \exp\left[-\frac{1}{2}\left[\alpha \frac{n}{L_1/2}\right]^2\right],\ k = 0, 1, \ldots, \frac{L_1 - 1}{2} \tag{6}$$

Similarly, in the stopband, the sample locations are

$$\omega_k = \omega_s + (\pi - \omega_s) \exp\left[-\frac{1}{2}\left[\alpha \frac{n}{L_2/2}\right]^2\right],\ k = 0, 1, \ldots, \frac{L_2 - 1}{2} \tag{7}$$

Here, choose the values of L_1 and L_2 are to be odd such that $\frac{L_1-1}{2} + \frac{L_2-1}{2} + 2 = \frac{N}{2}$. The generated samples are symmetric so only half of the samples used in design with two additional samples at 0 and $\omega_p - 0.01$.

Thus, steps in the proposed nonuniform frequency sampling method for FIR filter design are as follows

1. Given filter specifications are ω_p, ω_s, δ_p and δ_s.
2. Choose optimum order N of filter using Kaiser formula [1].
3. Generate the nonuniform sample separately in passband and stopband using Gaussian window function. Let's samples L_1 in passband and L_2 samples in stopband.
4. Set the desired frequency response values at given nonuniform samples. (Unity in passband and zero in stopband).
5. Generate NDFT matrix from nonuniform frequency samples.
6. Find the impulse response of filter.

5 Results and Discussions

The result of proposed method is compared with uniform frequency sampling (UFS) and Parks–McClellan (PM) method for low-pass filter design and discussed here. In the proposed method, nonuniform samples are generated separately in passband and stopband using Gaussian function from Eq. (2). The number of samples generated in passband as L_1 and in stopband as L_2 such that $L_1 + L_2 + 2 = N$. Considering such nonuniform samples, the low-pass filter is designed.

5.1 Generic Examples

Here, the low-pass filter is designed by four different specifications and its frequency response is studied. Consider the following cases for the filter design

1. LPF with $\omega_p = 0.1\pi$ and $\omega_s = 0.2\pi$
2. LPF with $\omega_p = 0.2\pi$ and $\omega_s = 0.3\pi$
3. LPF with $\omega_p = 0.5\pi$ and $\omega_s = 0.6\pi$
4. LPF with $\omega_p = 0.7\pi$ and $\omega_s = 0.8\pi$.

Assume $\delta_p = \delta_s = 0.01$ and the optimum order of the filter is calculated by Kaiser formula as $N = 37$. The following Fig. 1a–d shows the frequency response comparison for different specifications.

5.2 Practical Example: Filter for EEG Analysis

Consider the example of EEG signal analysis where the extraction of signal of interest from a noisy background is required. EEG signal analysis is helpful in various clinical applications like predicting epileptic seizer, classifying sleeping stages, etc. In EEG signal analysis the most important frequency range of interest is 0–30 Hz. So, the digital low-pass filter is used to limit the frequency band during preprocessing of EEG analysis. Digital filtering in EEG is generally done by FIR filter design methods. So, the proposed method is used to design digital LPF for EEG analysis.

In EEG analysis, the sampling rate is used 200 Hz and the passband frequency is $f_p = 30\,\text{Hz}$ and stopband frequency is $f_s = 40\,\text{Hz}$. Assume $\delta_p = \delta_s = 0.01$ and the optimum order of the filter is calculated by Kaiser formula as $N = 37$. Choose the value of $L_1 = 11$ and $L_2 = 25$ and two additional samples in passband at 0 and $\omega_p - 0.01$ such that $L_1 + L_2 + 2 = N$. The following Fig. 2 shows the frequency response comparison using different methods of FIR filter design.

From the Tables 1 and 2, it is observed that the investigated method has advantage over uniform frequency sampling in passband ripple and stopband ripple. In comparison with Parks–McClellan method, overall frequency response is approximately the same.

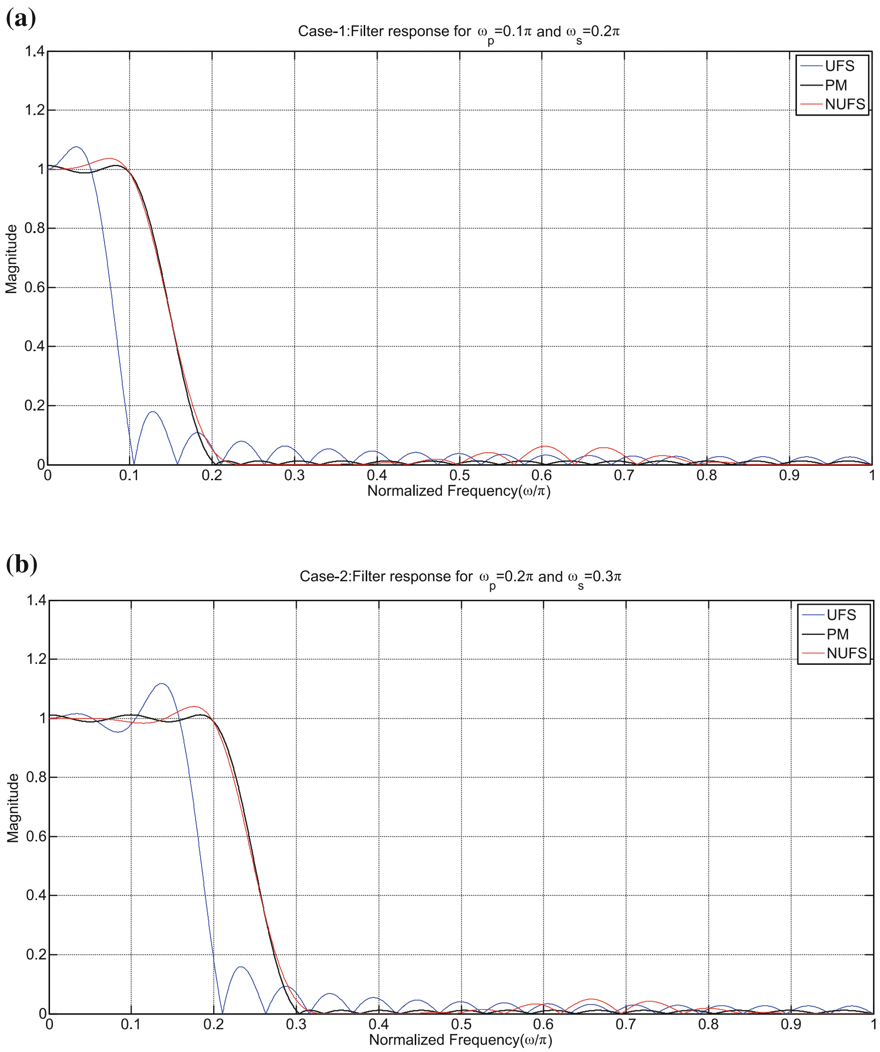
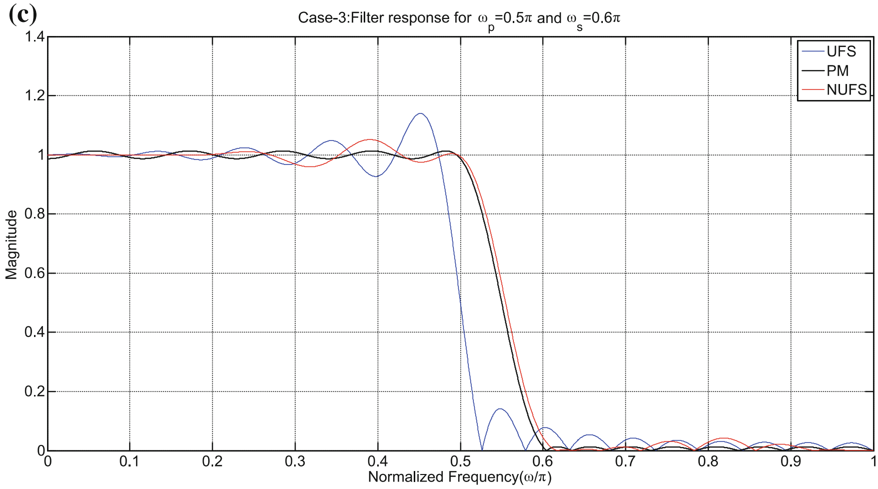


Fig. 1 Frequency response of filter for different cases

(c)



(d)

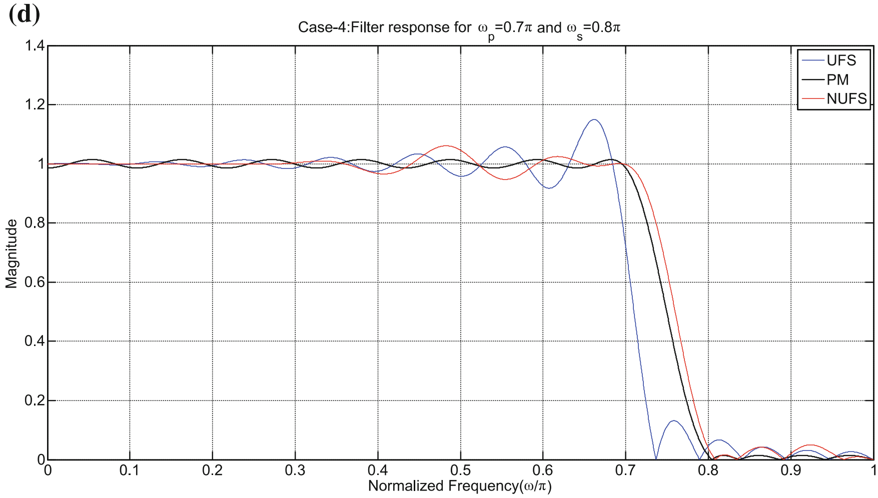


Fig. 1 (continued)

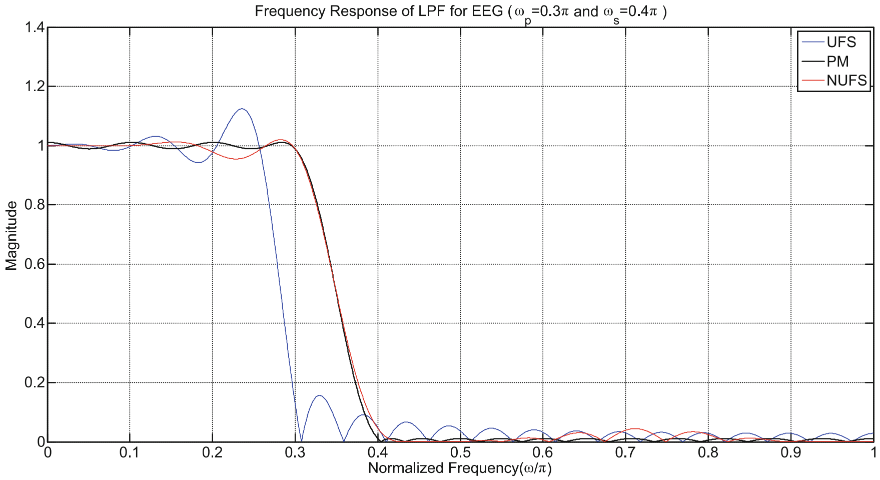


Fig. 2 Frequency response of EEG filter using different methods

Table 1 Performance analysis for different cases

Cases	Methodology	Realized values of		
		δ_p	δ_s	ω_c/π
1	UFS	0.0758	0.1801	0.0716
	PM	0.0124	0.0124	0.1344
	NUFS	0.0365	0.0625	0.1328
2	UFS	0.1184	0.1594	0.1754
	PM	0.0115	0.0116	0.2346
	NUFS	0.0398	0.0494	0.2318
3	UFS	0.1401	0.1421	0.4903
	PM	0.0131	0.0131	0.5345
	NUFS	0.0516	0.0426	0.539
4	UFS	0.1495	0.1318	0.7004
	PM	0.0143	0.0143	0.7338
	NUFS	0.0609	0.0501	0.7447

Table 2 Performance analysis for EEG filter

Methodology	Realized values of		
	δ_p	δ_s	ω_c/π
UFS	0.1296	0.1514	0.2802
PM	0.0116	0.0117	0.3346
NUFS	0.0451	0.0445	0.3337

6 Conclusions

In this paper, a new approach is investigated for the design of FIR filter based on nonuniform frequency sampling. This method is simple and easy to design. This method uses the NDFT which is nonuniform sampling of Z-transform along unit circle. Here, the Gaussian window based nonuniform samples is used to sample ideal frequency response of low-pass filter. The filter coefficients are determined by taking inverse NDFT of sampled frequency response. It is found that the frequency response of filter by the proposed method is better than uniform frequency sampling method in passband and stopband for the optimum order of filter. And also matches with the frequency response of filter using well-known Parks–McClellan algorithm.

References

1. Mitra, S.: Digital Signal Processing—A Computer Based Approach, Fourth edn. McGraw Hill Education (India) Private Limited, New Delhi (2013)
2. Proakis, J., Manolakis, D.: Digital Signal Processing-Principles, Algorithms, and Applications, Fourth edn. Pearson Publications (2007)
3. Rabiner, L., Gold, B., McGonegal, C.: An approach to the approximation problem for nonrecursive digital filters. IEEE Trans. Audio Electroacoust. **18**(2), 83–106 (1970)
4. Angelidis, E., Diamessis, J.: A novel method for designing FIR digital filters with nonuniform frequency samples. IEEE Trans. Signal Process. **42**(2), 259–267 (1994)
5. Burrus, C., Soewito, A., Gopinath, R.: Least squared error FIR filter design with transition bands. IEEE Trans. Signal Process. **40**(6), 1327–1340 (1992)
6. Bagchi, S., Mitra, S.: The Nonuniform Discrete Fourier Transform and Its Applications in Signal Processing. Springer Science+Business Media, LLC, New York (1999)
7. Bagchi, S., Mitra, S.: The nonuniform discrete fourier transform and its application in filter design: Part I - 1-D. IEEE Trans. Circuits Syst. II: Analog Digit. Signal Process. **43**(6), 422–433 (1996)
8. Marvasti, F.: Nonuniform Sampling-Theory and Practice. Springer Science+Business Media, LLC, New York (2001)
9. Yen, J.: On nonuniform sampling of bandwidth-limited signals. IRE Trans. Circuit Theory **3**(4), 251–257 (1956)
10. Neagoe, V.: Chebyshev nonuniform sampling cascaded with the discrete cosine transform for optimum interpolation. IEEE Trans. Acoust. Speech Signal Process. **38**(10), 1812–1815 (1990)
11. Tzvetkov, K., Tarczynski, A.: Digital filtering of band-limited signals using periodic nonuniform sampling. In: 2nd International Conference on Signals, Circuits and Systems, Monastir, pp. 1–6 (2008)
12. Margolis, E., Eldar, Y.: Nonuniform sampling of periodic bandlimited signals. IEEE Trans. Signal Process. **56**(7), 2728–2745 (2008)
13. Harris, F.: On the use of windows for harmonic analysis with the discrete Fourier transform. Proc. IEEE **66**(1), 51–83 (1978)

Detection of Epileptic Seizure Using Wavelet Transform and Neural Network Classifier

S. M. Wani, S. Sabut and S. L. Nalbalwar

Abstract The electroencephalograph (EEG) signals are most widely used for identification of neurological diseases like epilepsy, Alzheimer's, and other brain diseases. Detection of epileptic activity requires a detailed analysis of the entire length of the EEG data. In this paper, we proposed an automated detection of epileptic seizure using energy distribution of wavelet coefficient in each sub-band frequencies of the EEG signals. The performance of the proposed method is investigated using signals obtained from public EEG database at the University Hospital Bonn, Germany. Initially, the EEG signals are de-noised and decomposed into sub-bands using discrete wavelet transform (DWT), Then wavelet energy distribution in each sub-band is calculated and used as a feature set. Finally, artificial neural network (ANN) used to classify the feature set with ANN. The method was tested on EEG data sets obtained from that belongs to three subject groups: (a) healthy, (b) seizure-free interval, and (c) epileptic syndrome during a seizure. The test result shows that the proposed method for detecting epileptic seizure can achieve an overall classification accuracy of 95%. The proposed method can be used efficiently for recognition of epileptic seizures.

Keywords Epilepsy · EEG signal · Discrete wavelet transform · Energy distribution · Neural network classifier

S. M. Wani (✉) · S. Sabut
Department of Electronics Engineering, Ramrao Adik Institute of Technology, Navi Mumbai, India
e-mail: wsayali123@gmail.com

S. Sabut
e-mail: sukanta207@gmail.com

S. L. Nalbalwar
Department of Electronics & Telecommunication Engineering, Dr. Babasaheb Ambedkar Technological University, Lonere, India

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_75

1 Introduction

Epilepsy is one of the major neurological brain disorders which commonly occur due to one or more chronic conditions. Epilepsy is characterized by recurrent and uncontrollable electrical seizures due to sudden alteration in the electrical activity of the brain in the cortex [1]. EEG is a valuable tool to monitor the nonlinear electrical function of the brain activity. Epileptic seizures are commonly detected and tested by observation of EEG signal [2]. Currently, in the most cases, diagnosis of epileptic seizures is done manually through the observation of EEG signals which includes spikes, sharp waves, and spike-and-wave complexes not only during a seizure but also a short time before and between seizures [3]. An area of great interest is the development of efficient and reliable techniques for automatic diagnosis of early onset of seizures or even predicting before physical manifestations begin [4]. The DWT has been used effectively for detecting the epileptic seizer [5]. In many recent works, the automatic detection and decision-making system using neural network and support vector machines classifier has been proposed to recognize the epileptic seizures [6–12]. A recent review presented that in multi-paradigm approach by integrating wavelet transform and neural networks is the most effective method for automated EEG-based diagnosis of epilepsy [13].

This paper presents an automatic technique to detect the epileptic seizure in EEG signals using discrete wavelet transform with multi-resolution analysis. Specifically, five frequency sub-bands decomposition is obtained by applying DWT on EEG signal. Furthermore, the wavelet energy distribution at each sub-band levels is the most significant parameter to identify epileptic seizures is extracted as feature to make a feature set. The feature set is used as input to the NN classifier to classify three types of epileptic seizures.

2 Methodology

The workflow of the proposed method is shown in Fig. 1. The EEG signals are nonstationary time series signals that provides only the information in the time or the frequency domain. The wavelet transforms which provides an efficient alternative method that circumvents the disadvantages of the Fourier transform as far as the analysis of nonstationary signals is concerned.

2.1 Materials

In this work, the DWT method is employed for decomposition of signal into five different EEG components (δ, θ, α, β, and γ). Specifically, the fourth-order Daubechies (db4) wavelet is selected due to its good local approximated performance for nonsta-

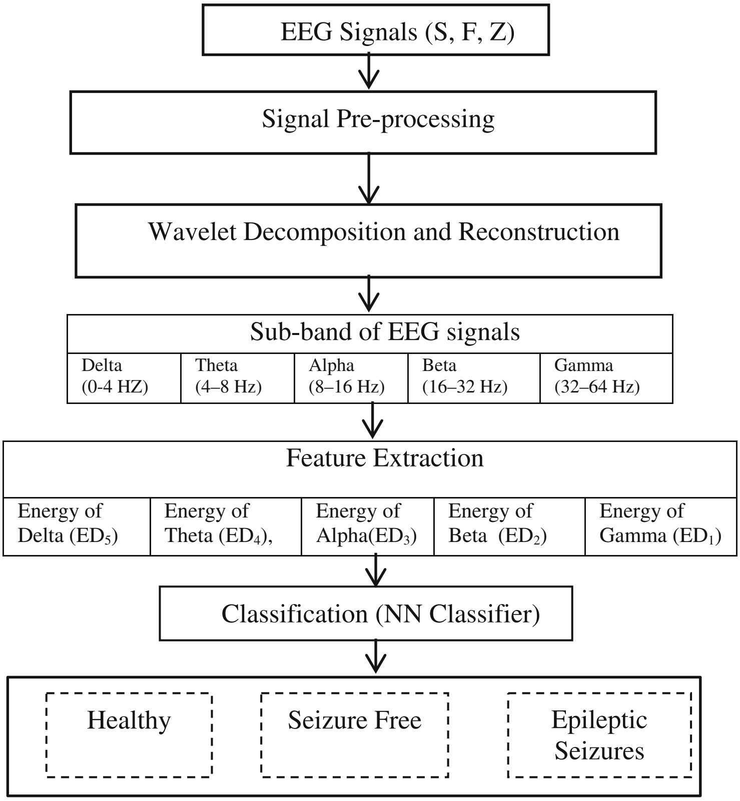


Fig. 1 Block diagram of the proposed epileptic classification framework

tionary signals [2]. Five frequency sub-bands (Delta (0–4 Hz), Theta (4–8 Hz), Alpha (8–16 Hz), Beta (16–32 Hz), and gamma (32–64 Hz) of clinical interest are obtained using the wavelet decomposition and at different levels. The wavelet energy distribution at each resolution levels are extracted as a features because of its good localizing properties, followed by a well-known ANN classifier is used to identify the epileptic seizures from EEG signals. A detailed flowchart of our proposed classification framework is shown in Fig. 1. The EEG signals used in this paper are obtained from public database at the University Hospital Bonn, Germany. The dataset recording is done with 128-channel amplifier system. After 12-bit analog-to-digital conversion, the data containing 100 single channel segments having sample points N = 4096 samples were written continuously onto data acquisition system at a sampling rate of 173.61 Hz. However, that time series have 0.5–85 Hz spectral bandwidth of the acquisition system. The EEG database consists of three subsets (Z, S, and F), the corresponding Nyquist frequency bandwidth is 86.8 Hz for 23.6-s duration. Z is obtained from five healthy volunteers through surface electrodes for open eye conditions while datasets S, F is acquired from epileptic patients. The datasets F are acquired during seizure-free intervals, while the dataset S only contains the seizure activity [11].

2.1.1 Discrete Wavelet Transform (DWT)

The WT is a spectral estimation technique in which any general function can be expressed as an infinite series of wavelets. The basic idea underlying wavelet analysis consists of expressing a signal as a linear combination of a particular set of functions (wavelet transform, WT), obtained by shifting and dilating one single function called a mother wavelet. The decomposition of the signal leads to a set of coefficients called wavelet coefficients [3]. The key feature of wavelets is the time–frequency localization. It means that most of the energy of the wavelet is restricted to a finite time interval. It means that most of the energy of the wavelet is restricted to a finite time interval as shown in Fig. 2.

Multi-resolution decomposition of time-domain signal gives approximation (A1) and detail information (D1) of signal at different scales which is simply obtained by correlating the original signal with wavelet function of different sizes. At the first level of decomposition, given a time-domain signal $x(n)$ of length n passes through half band high-pass (h(n)) and half band low-pass filter (g(n)) in which signal is convolved with impulse response of filter. The convolution operation in mathematical term is as follows:

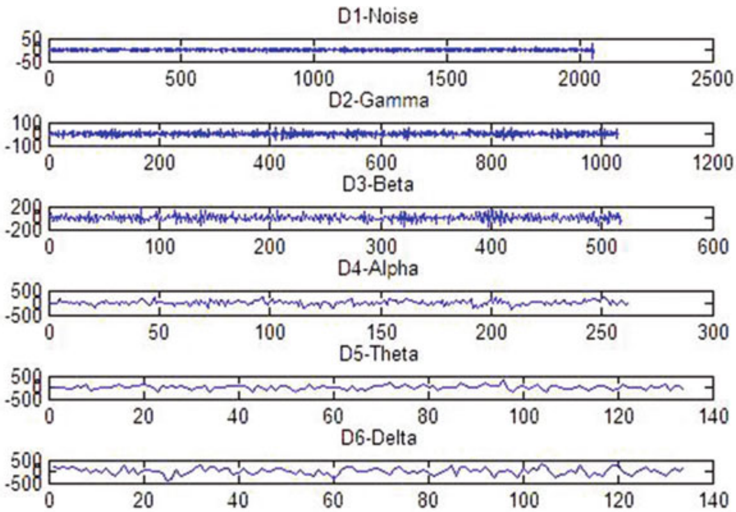


Fig. 2 Approximation and detailed coefficients of db 4 on an EEG signal

$$x(n) * h(n) = \sum_{k=\infty}^{-\infty} x(k) \cdot h(n-k) \tag{1}$$

In wavelet Transform subsampling by 2, reduces half number samples which double scale of signal. The output after subsampling by 2 gives approximation (A1) and detail information (D1). Similarly, after splitting the approximation coefficients A1 into two parts using the same scheme, produces A2 and D2 and so on. So that decomposition of signals into several frequency sub-bands requires a selection of an appropriate number of decomposition levels and wavelet function.

The DWT employs two sets of functions called scaling functions and wavelet functions, which are related to low-pass and high-pass filters, respectively. The decomposition of the signal into the different frequency bands is merely obtained by consecutive high-pass and low-pass filtering of the time-domain signal. All wavelet transforms can be specified in terms of a low-pass filter g, which satisfies the standard quadrature mirror filter condition

$$G(Z)G\left(Z^{-1}\right) + G(-Z)G\left(Z^{-1}\right) = 1 \tag{2}$$

where G(z) denotes the z-transform of the filter g. Its complementary high-pass filter can be defined as

$$H(Z) = ZG\left(-Z^{-1}\right) \tag{3}$$

$$G_{i+1}(Z) = G\left(Z^{I^2}\right)G_i(Z) \tag{4}$$

$$H_{i+1} = H(Z^{i^2})G_i(Z), \quad i = 0, \ldots, I-1 \tag{5}$$

with the initial condition G0(z) = 1. A two-scale relation in time domain is

$$\boldsymbol{g}_{i+1}(k) = [\boldsymbol{g}]_{\uparrow 2^i}\, \boldsymbol{g}_i(k) \tag{6}$$

$$h_{i+1}(k) = [h]_{\uparrow 2^i} \boldsymbol{g}_i(k) \tag{7}$$

where the subscript $[.]_{\uparrow m}$ indicates the up-sampling by a factor of m and k is the equally sampled discrete time. The normalized wavelet and scale basis functions can be defined as $\varphi_{i,l}(k)$,

$$\varphi_{i,l}(k) = 2^{i/2}\boldsymbol{g}_i\left(k - 2^i l\right) \tag{8}$$

$$\Psi_{i,l}(k) = 2^{i/2}\, h_i\left(k - 2^i l\right) \tag{9}$$

where the factor $2^{i/2}$ is an inner product normalization, i and l are the scale parameter and the translation parameter, respectively. The DWT decomposition can be described as

$$a_i(l) = x(k)^* \, \varphi_{i,l}(k) \tag{10}$$

$$d_i(l) = x(k)^* \, \Psi_{i,l}(k) \tag{11}$$

where $a_i(l)$ and $d_i(l)$ are the approximation coefficients and the detail coefficients at resolution *i,* respectively.

2.2 Wavelet Energy Distribution

The distorted signal is partitioned at different resolution levels and its energy is obtained using Parseval's theorem, and is mathematically represented as

$$ED_i = \sum_{j=1}^{N} \left|D_{ij}\right|^2 \; i = 1, \ldots, l \tag{12}$$

$$EA_i = \sum_{j=1}^{N} \left|A_{ij}\right|^2 \; i = 1, \ldots, l \tag{13}$$

where i = 1, …, l is the wavelet decomposition level from level 1 to N. N is the number of the coefficients of detail or approximate at each decomposition label. ED_i is the energy of the detail at decomposition level i and EA_i is the energy of the approximate at decomposition level l. Using wavelet decomposition with MRA, the 3 sets of the EEG signals are classified (20 EEG signals of the healthy patient, 20 EEG signals of the epilepsy patient in steady state, and 20 EEG signals of the epilepsy patient during the seizure) based on energy distribution in each resolution level. Energy distribution diagrams of EEG signals for different each set of EEG are shown in Fig. 3.

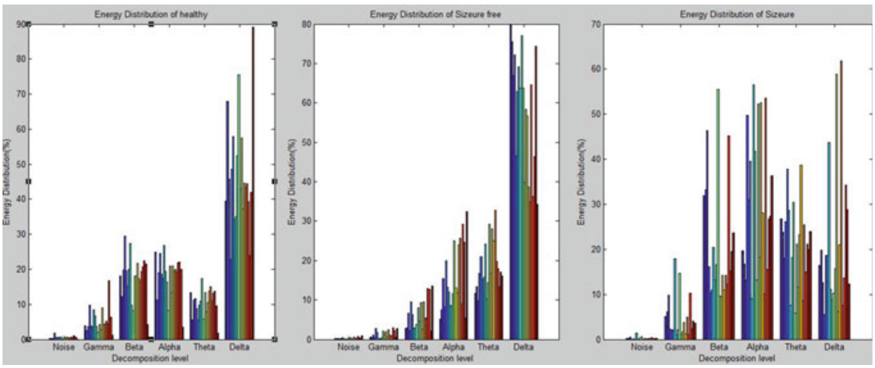


Fig. 3 Energy distribution diagram (%) **a** z set-20 EEG signals of healthy patient, **b** f set-20 EEG signals of epilepsy patient **c** s set-20 EEG signals of epilepsy patient with seizure

2.3 Neural Network (NN) Classifier

NNs are widely used in the biomedical signal processing field for complex pattern recognition and classification. Recent applications of NN in the signal processing and classification for biomedical problems can be found in several studies. The ANN is biologically inspired method for computing and classification of signal. This is an adaptive, nonlinear system that learns to perform mapping function on an input/output data in training phase. After the training phase, the ANN parameters are fixed and the system is deployed to solve the problem at the testing phase. The ANN is built with a systematic step-by-step procedure to optimize a performance criterion or to follow some implicit internal constraint, which is commonly referred to as the learning rule [14]. ANNs consists of input node attributes, one or more hidden layers, and one or more output nodes representing the output class. The data fed to input nodes as feature vector of variables and this information is passed to first hidden layer associated weights. The output of each hidden node are calculated as follows:

$$v_k = \sum_{i=1}^{n} w_{ki} x_{ki} \tag{14}$$

and

$$y_k = \sum_{i=1}^{n} \emptyset(v_k + v_{ko}) \tag{15}$$

where $x_1 \ldots x_n$ are input features, $w_{k1} \ldots w_{kn}$ are the connected unit of k, weighted, v_k is the net input, y_k is the output class, and Ø is the activation function of the neurons is sigmoid which is defined by

$$f_x = \frac{1}{1 + e^{-x}} \tag{16}$$

We used Levenberg–Marquardt backpropagation algorithm for feature classification. The energy distribution of each wavelet coefficients are extracted and used as training and testing data. The size of data set was 5×60. The input vector is applied as input to the three-layer ANN structure having five hidden neurons with TANGENT SIGMOID as activation function of neuron.

3 Results and Discussions

The levels are chosen such that those parts of the signal that correlates well with the frequencies of EEG signal that retained in the wavelet coefficients. The smoothing feature of Daubechies 4 (db4) is suitable for detecting changes of the EEG signals,

Table 1 Confusion table of classification results on the testing data set

Class	Healthy	Seizure free	Seizure	Accuracy
Healthy	19	1	0	95.0
Seizure free	1	18	0	94.7
Seizure	0	1	20	95.2
		Overall success rate		95.0

so we have chosen db4 as wavelet function with five levels of decomposition. We considered energy activity component D3, D4, and D5 which are dominant of signals and other components are discarded. Frequency bands corresponding to five decomposition levels for wavelet db4 with sampling frequency of 173.6 Hz of EEG signal are decomposed into details D1–D5 and one final approximation A5, as shown in Fig. 2.

It has been observed that energy distribution of gamma wave in healthy patients is about 6%, whereas in epilepsy, EEG signal energy distribution is much lower approximately 3%. Similarly, beta and alpha wave energy is quite similar with approximately 20%. In comparison with healthy patient data, the energy distribution of alpha and beta is much lower in seizure data. In the theta wave frequency, energy activity slightly lower with value around 10%. While in case of seizures, the energy distribution of theta and delta signal is much larger. Energy distribution of EEG signals during epileptic syndrome is significantly different from the first two cases as shown in Fig. 3. The feed-forward neural network was trained using the backpropagation algorithm. In the predicted classification, 20 signals of each case are considered. For example, for the testing data set of all EEG signal, it is found that out of total 60 signals 19 have been classified as healthy, 18 as seizure free, and 20 as epileptic syndrome during seizure. Since the majority of the predicted classification is in the Epileptic seizure. Table 1 presented are classification results of WNN algorithm where 20 data sets were used to train the ANN model and 5 data sets were used for testing process. The system can correctly classify 47 of the 60 different EEG signals in the testing set, is presented in Table 1. The proposed process achieved an accuracy of 95.0% in classifying epileptic seizure, which is better than few published articles such as Atoufi et al. [15] where the average accuracy was 60% and Ong et al. [16] where accuracy was 94.06%.

4 Conclusion

In this paper, we proposed an automatic method for detecting epileptic seizure using nonstationary EEG signal with wavelet decomposition and classified with NN classifier. The energy distribution of various EEG signal components is very important for detection and classification of epilepsy. It has been observed that use of energy distribution from EEG recordings obtained from subjects with epileptic seizure and normal subject's component using DWT enables classification, by a backpropagation algorithm in ANN, with a high degree of accuracy about 95%. Furthermore, it can reduce memory space, shorten preprocessing needs, and increase computation speed for the classification of an EEG signal. Hence, the proposed method could effectively be used for detection of seizure that could assist the physicians in diagnosis process.

References

1. Thakor, N.V., Tong, S.: Quantitative EEG Analysis Methods and Clinical Applications, Artech House, Boston/London, pp. 193–224 (2009)
2. Alotaiby, T., El-Samie, F.A., Alshebeili, S.A., Ahmad, I.: A review of channel selection algorithms for EEG signal processing. EURASIP J. Adv. Sig. Process. 66 (2015)
3. Shaker, M.M.: EEG wave classifier using wavelet transform and Fourier transform. Int. J. Biol. Life Sci. **1**, 85–90 (2005)
4. Omerhodzic, I., Causevic, E., Dizdarevic, K., Avdakovic, S., Music, M., Kusljugic, M., Hajdarpasic, E., Kadic, N.: First neurosurgical experience with the wavelet based EEG in diagnostic of concussion. In: Proceedings of 11th Congress of Neurosurgeons of Serbia, Serbia (2008)
5. Tzallas, A.T., Tsipouras, M.G., Fotiadis, D.I.: Automatic seizure detection based on time-frequency, analysis and artificial neural networks. Comput. Intell. Neurosci. 1–13 (2007)
6. Sharma, M., Pachorib, R.B., Acharya, U.R.: A new approach to characterize epileptic seizures using analytic time-frequency flexible wavelet transform and fractal dimension. Pattern Recogn. Lett. **94**, 172–179 (2017)
7. Wang, L., Xue, W., Li, Y., Luo, M., Huang, J., Cui, W., Huang, C.: Automatic epileptic seizure detection in EEG signals using multi-domain feature extraction and nonlinear analysis. Entropy **19**, 222 (2017)
8. Gandhi, T., Panigrahi, B.K., Anand, S.: A comparative study of wavelet families for EEG signal classification. Neurocomput. **74**, 3051–3057 (2011)
9. Patnaik, L.M., Manyam, O.K.: Epileptic EEG detection using neural networks and post-classification. Comput. Methods Progr. Biomed. **91**(2), 100–109 (2008)
10. Subasi, A., Ercelebi, E.: Classification of EEG signals using neural network and logistic regression. Comput. Methods Progr. Biomed. **78**(2), 87–99 (2005)
11. Andrzejak, R.G., Lehnertz, K., Rieke, C., Mormann, F., David, P., Elger, C.E.: Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. Phys. Rev. E **64**, 061907 (2001)
12. Zandi, A.S., Dumont, G.A., Javidan, M., Tafreshi, R., MacLeod, B.A., Ries, C.R., Puil, E.A.: A novel wavelet-based index to detect epileptic seizures using scalp EEG signals. In: Proceedings of IEEE Engineering in Medicine and Biology Society, pp. 919–922 (2008)
13. Faust, O., Acharya, U.R., Adeli, H., Adeli, A.: Wavelet-based EEG processing for computer-aided seizure detection and epilepsy diagnosis. Seizure **26**, 56–64 (2015)
14. Lancashire, L.J., Lemetre, C., Ball, G.R.: An introduction to artificial neural networks in bioinformatics-application to complex microarray and mass spectrometry datasets in cancer studies. Brief. Bioinform. **10**, 315–329 (2009)
15. Atoufi, B., Lucas, C., Zakerolhosseini, A.: A survey of multi-channel prediction of EEG signal in different EEG state: normal, pre-seizure, and seizure. In: Proceedings of the 7th International Conference on Computer Science and Information Technologies, Yerevan, Armenia (2009)
16. Lan, T., Erdogmus, D., Adami, A., Mathan, S., Pavel, M.: Channel selection and feature projection for cognitive load estimation using ambulatory EEG. Comput. Intell. Neurosci. 1–12 (2007)

Comparative Analysis of ICA, PCA-Based EASI and Wavelet-Based Unsupervised Denoising for EEG Signals

Ankita Bhatnagar, Krushna Gupta, Utkarsh Pandharkar, Ramchandra Manthalkar and Narendra Jadhav

Abstract Electroencephalography (EEG) can be used to study various brain activities related to human responses and disorders. EEG signal is prone to noises which are caused due to eye movements, power-line interference, muscle movements, etc. Therefore, to obtain refined EEG signals for further processing, it should be denoised. There are several methods by which EEG signals can be denoised, among which we have used Independent Component Analysis (ICA), Principal Component Analysis (PCA)-based Equivariant Adaptive Separation by Independence (EASI), and Wavelet-based unsupervised denoising methods. The performance of these methods is compared using Signal-to-Noise Ratio (SNR) and Percentage Root-mean-square Difference (PRD).

Keywords EEG · Denoising · ICA · PCA-based EASI · Wavelet

A. Bhatnagar · K. Gupta (✉) · U. Pandharkar · R. Manthalkar · N. Jadhav
Shri Guru Gobind Singhji Institute of Engineering and Technology, Vishnupuri, Nanded, Maharashtra, India
e-mail: guptakrushna@sggs.ac.in

A. Bhatnagar
e-mail: bhatnagarankita@sggs.ac.in

U. Pandharkar
e-mail: pandharkarutkarsh@sggs.ac.in

R. Manthalkar
e-mail: rrmanthalkar@sggs.ac.in

N. Jadhav
e-mail: jadhavnarendra@sggs.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_76

1 Introduction

EEG is a noninvasive technique of recording brain's spontaneous electrical activities generated by neurons. The quality of recorded EEG signals depends on various factors such as external noises like power-line interference and artifacts (muscle movements, eye movements, etc.). For analysis and processing of EEG signals, it is necessary to eliminate such noises to obtain more accurate and appropriate results. The corrupted EEG signal is, hence, preprocessed using denoising techniques which increases the reliability of EEG data. For removal of artifacts various denoising techniques can be used, among which ICA, PCA-based EASI, and wavelet denoising are prominently selected for this study.

Cardoso and Laheld [1] have given the algorithm for blind source separation known as EASI. The components obtained after PCA are used for extraction of sources using EASI algorithm. An unsupervised algorithm for artifactual component identification is proposed by Mahajan and Morshed [2], which is used here for components obtained after ICA and EASI. Hazra and Guhathakurta [3] measured the performance of different denoising algorithms and is measured using SNR and Mean Square Error (MSE). EEG signal artifacts (ocular) are normally in low-frequency region, hence, denoising is applied on low-frequency Senthil Kumar et al. [4] implemented a method to remove ocular artifacts from EEG data using wavelet transform without an EOG reference channel. Most studies have done analysis on different techniques for denoising EEG signal but not many have compared ICA, PCA-based EASI method, and Wavelet denoising.

2 Materials and Methods

In this work, the EEG data taken into consideration has been acquired by Jadhav et al. [5] using wireless EMOTIV EPOC+ with 14 electrodes (placed according to standard 10–20 system) and sampling frequency of 128 Hz. The protocol shown to subjects has a part of meditation initially followed by slides of pictures depicting four types of emotion. The experimental protocol is given in Fig. 1. Images (I1, I2, I3, and I4) depicting emotions happy, sad, angry, and relax are shown with blank spaces in between.

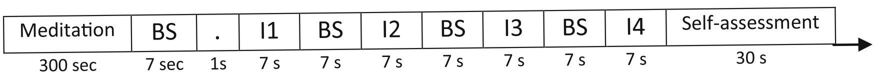


Fig. 1 Protocol

2.1 Independent Component Analysis (ICA)

EEG signals follow all the assumptions of ICA model and use of ICA model on EEG has been validated [6]. If we have n different observed signals, namely x_1, x_2, x_3, … x_n and some random variables as s_1, s_2, s_3, … s_n. These observed signals can be expressed in linear combination as [7]

$$x_i = a_{i1}s_1 + a_{i2}s_2 + a_{i3}s_3 \ldots a_{in}s_n \tag{1}$$

For i = 1, 2, 3 … n where, a_{ij} are real constants, where j = 1, 2, 3… n.

ICA technique is used to estimate the random components $s_1, s_2, s_3, s_4, \ldots, s_n$ which are also known as the independent components (ICs). In vector-matrix notation, this mixing model can be expressed as

$$x = As \tag{2}$$

To estimate sources, a separation matrix W should be found. Based on experimental results on statistical and computational load analysis of various ICA algorithms, fixed-point Fast-ICA algorithm using *tanh* nonlinearity with symmetrical orthogonalization is chosen for finding ICs [7].

Fast-ICA implementation for finding Independent Components.
Preprocessing. For better conditioning and making IC estimation simpler, preprocessing is done by centering and whitening [7] on the acquired EEG data x.

Centering—Zero-mean data $\bar{x}$ is computed as follows:

$$\bar{x} = x - Ex \tag{3}$$

Whitening—Eigen Value Decomposition (EVD) of covariance matrix of $\bar{x}$ is used to compute whitened data z

$$z = ED^{-\frac{1}{2}}E^T\bar{x} \tag{4}$$

where E is orthogonal matrix of eigenvectors, D is diagonal matrix of its eigenvalues.
Fast-ICA algorithm. Algorithm of fixed-point fast-ICA for several units using symmetrical orthogonalization is given by Hyvärinen and Oja [8] using which separation matrix W such that $\tilde{s} = Wz$ is computed. Here, $\tilde{s}$ is the matrix of estimated sources or independent components. W is updated using the following equation:

$$w_i \leftarrow E\{zg(w_i^T z)\} - E\{g'(w_i^T z)\}w_i \tag{5}$$

where, $g(y) = tanh(y)$, $W = (w_1, w_2, w_3 \ldots, w_m)^T$ and m is the number of independent components to be estimated.

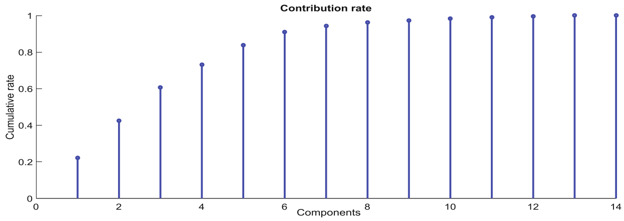


Fig. 2 Cumulative contribution plot

Artifactual component identification and elimination. After finding ICs using fast-ICA, for identification and elimination of artifactual components, kurtosis property of ICs and DWT is used as given in Sect. 2.3.

2.2 *Principal Component Analysis Based EASI*

In PCA algorithm, data dimension is reduced by getting component with maximum weight without losing more information. From principal components, i.e., compressed data, the individual components found using (EASI).

PCA algorithm. The N-channel EEG signal X_{MxN} is centered to $\bar{X}$. By applying rotated orthogonal coordinate system, the data is uncorrelated. Correlation matrix R_{xx} is obtained. The eigenvalues and eigenvectors are found. Eigenvalues are arranged such that $\lambda_1 > \lambda_2 > \ldots \lambda_N$ and accordingly eigenvectors are arranged $c_1, c_2 \ldots, c_N$. Where, $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_N]$ and $C = [c_1, c_2, \ldots, c_N]$. Equation 6 is used to obtain the matrix of principal components [7, 9]

$$Y = C^T\bar{X} \tag{6}$$

where $Y = [y_1, y_2 \ldots y_N]^T$. To select the number of principal components, the contribution rate of each component is to be considered using eigenvalues. The contribution rate of each component is found by $\frac{\lambda i}{\sum_{k=1}^{N} \lambda i}$.

The cumulative contribution rates are plotted in Fig. 2 and eliminated according to its weight. Components of cumulative contribution rate below 99.604 are selected to avoid unwanted data loss. Hence, 12 components among 14 are selected [9].

EASI algorithm. For Blind Source Separation (BSS), EASI is used. EASI algorithm is based on serial updation where transformation on data as well as parameter is equivalent. Assuming source signal to be s and x is matrix of selected principal components, then x can be given as x = As, where, A is the mixing matrix.

To estimate these sources, we find a separating matrix W such that

$$Y = W\,x \tag{7}$$

Here, y is very close to s. Ideally, y = s for obtaining accurate sources which implies $A = W^{-1}$. The separating matrix for EASI algorithm is updated as follows [9, 10]:

$$W(k+1) = W(k) - \mu\left[Y(k)\cdot Y^T(k) - I + g(Y(k))\cdot Y^T(k) + g\left(Y^T(k)\right)\cdot Y(k)\right]W(k) \tag{8}$$

where μ is learning rate, I is identity matrix, $W = [W_1, W_2, \ldots W_n]^T$ is separation matrix and $g(Y) = [g(y_1), g(y_2) \ldots, g(y_n)]^T$, g(.) is considered as any nonlinear function. The estimation of sources is given by Y = WX. If source is super-Gaussian, then nonlinear function $\tan^{\gamma}h$ ($^{\gamma} > 1$) is used [9]. Main interference in this experiment is EOG which is super-Gaussian in nature. Hence, nonlinear function used is *tan10 h*.

Artifactual component identification and elimination. Noise containing components are selected by kurtosis values of each component. Further, the thresholding is done using DWT (Discrete Wavelet Transform) instead of direct source elimination to avoid the data loss as given in Sect. 2.3.

2.3 Artifact Elimination for ICA and PCA-Based EASI

Kurtosis and confidence interval. Data having high kurtosis depict high tailed nature similar to ocular noise [2].

Procedure for finding artifactual components.

(a) Compute kurtosis for each component

$$\text{kurtosis} = \sum_{i=1}^{n}\left(\frac{\left(y_i - \bar{y}\right)/n}{s^{\,4}}\right) \tag{9}$$

where n is number of samples, s is standard deviation of n sample data y, and $\bar{y}$ is mean of samples.

(b) Find upper bound of 95% Confidence Interval (CI) [2]

$$CI = \bar{y} + \frac{s}{\sqrt{N}}t \tag{10}$$

where N is number of components, t is the t-value for specific percentage of confidence interval. By calculating CI, threshold value is set, above which the signals are considered to be artifactual.

Further, the artifactual components detected are thresholded using DWT.

Discrete Wavelet Transform (**DWT**). For denoising of signal, n level DWT is applied on signal. Levels are number of times that DWT is applied on approximate coefficient of previous level. Here, four-level mother wavelets with db4 kernel are applied on signal. Thresholding of DWT coefficient is done using hard thresholding. For thresholding, threshold value is defined using universal threshold technique [4].
Universal Threshold. It is a good approach for statistical smoothness whose asymptotic behavior is better the MSE [4]. If N is number of coefficient in series and X is series of wavelet coefficients, it is formulated as

$$\mathrm{th} = \sigma\sqrt{2\log N} \tag{11}$$

$$\sigma^2 = \frac{\mathrm{median}(X)}{0.6745} \tag{12}$$

After thresholding of coefficients, reconstruction of signal is carried out by applying inverse DWT on modified wavelet coefficients. That reconstructed signal is thresholded signal using DWT.

Reconstruction of ICA and PCA.

For ICA. After artifact selection using kurtosis and thresholding data using DWT, the obtained result is multiplied to inverse of separation matrix to obtain denoised signal. The equation is given by [8]

$$x_d = W^{-1}s_t \tag{13}$$

where x_d is the denoised signal, s_t is artifact free source signal, and W is the separation matrix of ICA.
For PCA. As noise-free components are obtained, further reconstructed principal components can be obtained by multiplying inverse of separation matrix inverse with thresholded component matrix.

$$Y_{re} = W^{-1}Y \tag{14}$$

where Y_{re} is reconstructed components, W is the separation matrix of EASI, and Y is the source after artifact removal.

Further reconstruction of noise-free signal can be obtained from reconstructed principal component matrix by multiplying it with weight matrix of principal components [9].

$$X_{re} = C^{-1}Y_{re} \tag{15}$$

where X_{re} is denoised signal, C is the PCA weight matrix.

2.4 Wavelet Denoising

Identification of Noise Region (**Method 1**). To identify the spike region, Stationary Wavelet Transform is used. SWT provides translational invariance which is important to identify random noise. As EEG data samples at 128 samples per sec (2^7), sixth (j − 1) level SWT is applied which gives detailed and approximate coefficient. The noise region is identified using the following method which automatically marks spikes in EEG data.

If detailed coefficient of sixth level exceeds 40% of maximum value of detailed coefficient present, then mark that as spike. Similarly, the whole spikes region is marked.

Thresholding Technique for method 1. In order to remove spikes from EEG data, thresholding is applied on it. This only removes ocular noise from data. Threshold is defined as

$$\mathrm{Th} = \mathrm{N} * \frac{\mathrm{x}' - \sigma}{\mathrm{x}' + \sigma} \tag{16}$$

where N is a positive integer, ranging from 100 to 150, x′—Mean of all samples, σ—Standard deviation of all samples.

From that spike, EEG data is separated and ocular noise data is removed using above threshold. The thresholding function that is used is as follows [4, 11]:

$$\mathrm{D_j} = (-0.7) * \mathrm{D_j} \quad \text{when } \mathrm{D_j} \geq \mathrm{Th} \tag{17}$$

$$\mathrm{D_j} = \mathrm{D_j} \quad \text{when } \mathrm{D_j} < Th \tag{18}$$

Regeneration of ocular noise-free EEG data. After thresholding of noise region in EEG data, it is necessary to regenerate noise-free EEG data. To regenerate this, inverse of stationary wavelet transform is applied on approximate and new series of detailed coefficient. By applying ISWT, ocular noise-free EEG data is obtained.

Method 2. After the applying above method, only spike region is eliminated. Now, the non-spiked noise must be removed for analysis further using DWT and thresholding method. Further, decomposition of four levels is calculated.

Thresholding Techniques for method 2. Threshold selection for the denoising is one of the important tasks [3, 12, 13].

Soft Thresholding. In soft thresholding, if coefficient value exceeds the threshold, then coefficient is modified or otherwise kept as it is.

$$Y(n) = \mathrm{sgn}(X'(n)) * i\left(\left|X'(n) - T\right|\right) \quad \text{when } X'(n) \geq Th \tag{19}$$

$$Y(n) = X'(n) \quad \text{when } X'(n) < Th \tag{20}$$

Stein Unbiased Risk Estimator. In statistics, Stein's unbiased risk estimate (SURE) is an unbiased estimator of the mean-squared error of a nearly arbitrary, nonlinear biased estimator.

$$\hat{R} = n\sigma^2 + \left|\left|X - \hat{\mu}\right|\right|^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\delta\,\hat{\mu}}{\delta X_i}(X) \tag{21}$$

where σ is standard deviation and $\hat{\mu}$ is mean of wavelet coefficient X of each level.

In regaining of signal, Inverse Wavelet Transform of modified wavelet coefficient is used. Applying IDWT on wavelet coefficient, results in denoised EEG signal. The signal obtained at output is not only an ocular free but also other noise free.

2.5 Performance Parameters

The performances of the denoising methods are checked by SNR and PRD [14].

$$SNR = 10\log\left\{\frac{\sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N} e_i^2}\right\} \tag{22}$$

$$PRD = 100\sqrt{\left\{\frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} x_i^2}\right\}} \tag{23}$$

where N is number of samples in signal x, e is the error (difference between original and denoised signal), and x is the original signa

Table 1 Kurtosis value of independent components by ICA

IC1	IC2	IC3	IC4	IC5	IC6	IC7	IC8	IC9	IC10	IC11	IC12	IC13	IC14
4.35	3.29	4.20	5.08	17.67	**19.73**	4.41	**22.62**	**22.21**	7.61	**24.91**	**23.58**	5.67	15.94

Table 2 Kurtosis of separated sources by PCA

S-1	S-2	S-3	S-4	S-5	S-6	S-7	S-8	S-9	S-10	S-11	S-12
11.14	12.77	13.70	10.77	**15.45**	**15.85**	4.42	12.24	9.84	6.81	10.04	**21.45**

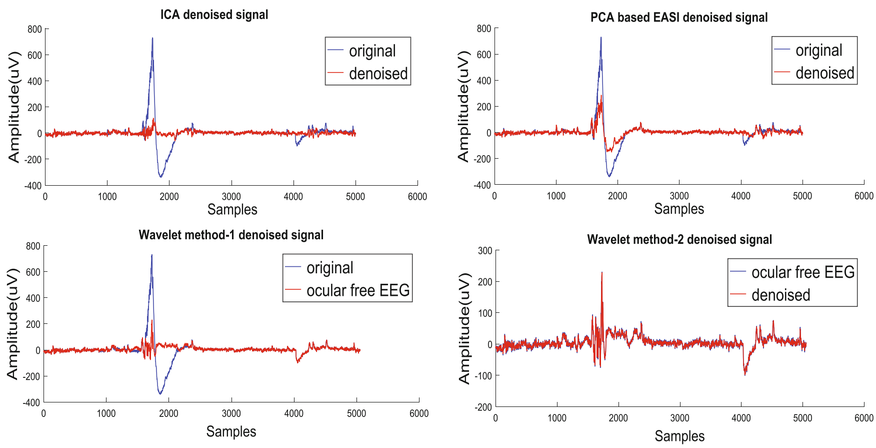


Fig. 3 Denoised signal by all methods

3 Results

The value of 95% CI over mean is 17.9118. Therefore, independent components IC6, IC8, IC9, IC11, and IC12 are selected as artifactual components.

The value of 95% CI over mean is 14.8248. Therefore, sources S-5, S-6, and S-12 are selected as artifactual components (Tables 1 and 2).

The average SNR values and PRD for ICA, PCA-based EASI, and wavelet denoising are 15.8844 ± 4.6606, 55.9983 ± 16.2655, 53.4595 ± 14.1662, and 24.2067 ± 8.5159, 1.6965 ± 0.9650, 8.9043 ± 4.6342, respectively.

Denoised signals by all the methods are given in Fig. 3 which shows elimination of noise. By Fig. 3 it can be observed that ocular noise elimination is more than others. Wavelet method 1 eliminates ocular noise and method 2 gives high-frequency noise-free data.

4 Conclusion

PCA-based EASI gives a better result than ICA and wavelet denoising when SNR and PRD are considered. But PCA-based EASI has a problem with convergence which does not give well-separated sources and can lead to data loss. Experimentally, to obtain well-separated sources, ICA is better method and it gives reliable output. As in wavelet denoising, the noise is removed only from selected segments, and it has better performance in all aspects.

References

1. Cardoso, J.-F., Laheld, B.H.: Equivariant adaptive source separation. IEEE Trans. Sig. Process. **44**(12) (1996)
2. Mahajan, R., Morshed, B.I.: Unsupervised eye blink artifact denoising of EEG data with modified multiscale sample entropy, kurtosis, and wavelet-ICA. IEEE J. Biomed. Health Inform. **19**(1) (2015)
3. Hazra, T.K., Guhathakurta, R.: Comparing wavelet and wavelet packet image denoising using thresholding techniques. Int. J. Sci. Res. (IJSR) **5**(6) (2016)
4. Senthil Kumar, P., Arumuganathan, R., Sivakumar, K., Vimal, C.: Removal of ocular artifacts in the EEG through wavelet transform without using an EOG reference channel. Int. J. Open Prob. Comput. Math. **1**(3) (2008)
5. Jadhav, N., Manthalkar, R., Joshi, Y.: Effect of meditation on emotional response: an EEG-based study. Biomed. Sig. Process. Control **34**, 101–113 (2017)
6. Makeig, S., Bell, A.J., Jung, T.-P., Sejnowski, T.J.: Independent component analysis of electroencephalographic data. In: Proceedings of Advances in Neural Information Processing Systems (NIPS 1995), vol. 8 (1995)
7. Hyv¨arinen, A., Karhunen, J., Oja, E.: Independent component analysis. Wiley (2001)
8. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Netw. **13**(4–5) (2000)
9. Dong Kang, F., Luo Zhizeng, S.: A method of denoising multi-channel EEG signals fast based on PCA and DEBSS Algorithm. 2012 International Conference on Computer Science and Electronics Engineering, (2012)
10. Simranpreet Kaur, F., Sheenam Malhotra, S.: Various Techniques for Denoising EEG signal: A Review. International Journal Of Engineering and Computer Science ISSN:2319-7242 Volume 3 Issue Page No. 7965-7973, (2014)
11. Tibshirani, R.: Stein's unbiased risk estimate. Statistical Machine Learning. Springer (2015)
12. Khatwani, P., Tiwari, A.: Removal of noise from EEG signals using cascaded filter—wavelet transforms method. Int. J. Adv. Res. Electr. Electron. Instrum. Eng. **3**(12) (2014)
13. Estrada, E., Nazeran, H., Sierra, G., Ebrahimi, F., Mikaeili, M.: Wavelet EEG denoising for automatic sleep stage classification (2011). https://www.researchgate.net/publication/221632560
14. Walters-Williams, J., Li, Y.: Using invariant translation to denoise electroencephalogram signals. Am. J. Appl. Sci. **8**(11), 1122–1130 (2011)
15. Princy, R., Thamarai, P., Karthik, B.: Denoising EEG signal using wavelet transform. Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) **4**(3) (2015)
16. Al-Qazzaz, N.K., Hamid Bin Mohd Ali, S., Ahmad, S.A., Islam, M.S., Escudero J.: Selection of mother wavelet functions for multi-channel EEG signal analysis during a working memory task. Sensors **15**, 29015–29035 (2015). https://doi.org/10.3390/s151129015

17. Garg, S., Narvey, R.: Denoising & feature extraction of EEG signal using wavelet transform. Int. J. Eng. Sci. Technol. **5**(6) (2013)
18. Zheng-you, H.E., Xiaoqing, C., Guoming, L.: Wavelet entropy measure definition and its application for transmission line fault detection and identification. In: International Conference on Power System Technology (2006)
19. Delorme, A., Sejnowski, T., Makeig, S.: Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. NeuroImage **34**(4) (2007)

Analyzing Effect of Meditation Using Higher Order Crossings and Functional Connectivity

Shruti Phutke, Narendra Jadhav, Ramchandra Manthalkar and Yashwant Joshi

Abstract People are experiencing difficulties in adapting to the rapid changes in work and social fabric due to the evolution of advanced technologies in everyday life. Health and well-being of an individual in the existing world is important for proper living. Meditation improves the adaptability of an individual to live a healthy and social life. To verify this, an experiment is designed with the simple meditation practice called Focused Attention for 8 weeks. The brain activity is recorded of 11 subjects using EMOTIV EPOC+ EEG device before (pre-meditation) and after (post-meditation) meditation. Features called Higher Order Crossings and Functional Connectivity are used to analyze the effect of meditation. The results indicated a decrease in HOC values for frontal, parietal, and occipital lobes and increase in HOC of temporal lobe. The interhemispheric connectivity increased after meditation practice.

Keywords Meditation · EEG · Higher order crossings · Functional connectivity

1 Introduction

Students in recent days experience stress while adapting the rapid changes in work and social life. Increased stress leads to improper attention towards work and studies. Meditation is a time-honored method to improve the adaptability of an individual for

S. Phutke (✉) · N. Jadhav · R. Manthalkar · Y. Joshi
Centre of Excellence in Signal and Image Processing, Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, India
e-mail: shrutiphutke@gmail.com

N. Jadhav
e-mail: jadhavnarendra@sggs.ac.in

R. Manthalkar
e-mail: rrmanthalkar@sggs.ac.in

Y. Joshi
e-mail: yvjoshi@sggs.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_77

healthy life. Training the human mind to observe breath makes it capable of acquiring the state of peace and attains well-being. Mindfulness is the process of remaining consciously aware of what is happening in the field of experience. Mindfulness Meditation (MM) is the way to meditate where one has to focus on the movement of the abdomen or on the incoming and outgoing breath. If the attention on the breath is distracted then the person should notice it passively and return his/her focus on breathing. The research is going on to analyze the effect meditation on brain. The brain analysis can be done using various methods such as fMRI, PET, EEG, etc. The Electroencephalogram (EEG) is a noninvasive method to acquire the neuronal activity from the electrodes placed on the brain scalp. As compared to other methods (fMRI, PET) [1], EEG gives high time resolution while capturing the data. The EEG band consists of different frequency bands such as delta (1–4 Hz), theta (4–8 Hz), alpha (8–16 Hz), beta (16–30 Hz), and gamma (30–100 Hz). The human brain consists of three parts cerebrum, cerebellum, and brain stem. The body movements and coordination is controlled by cerebellum. The cerebrum is again divided into four different lobes called as frontal lobe, parietal lobe, temporal lobe, and occipital lobe. Each brain lobe deals with different functions such as problem solving, emotions, hearing, etc. [2]. In this era of overloading information, it is essential to mitigate the stress and anxiety by various available means. In this paper, the Higher Order Crossings (HOC) and Functional Connectivity in pre- and post-meditation is analyzed. Rest of the paper is described as follows: Sect. 2 gives the related work done in the area of meditation analysis. Section 3 elaborates the Materials and Methods used for analysis. In Sect. 4, the result is discussed and Sect. 5 concludes the paper.

2 Related Work

Various studies have been done on Meditation using EEG. Pei Chen et al. reported microstate analysis for the spatiotemporal characteristics of Chan meditation [3]. Kang Ming Chang proposed a method of quantifying EEG and flash visual evoked potential for Zen Meditation [4]. The impact of meditation on emotional processing for meditation and control group using event-related potentials (ERP) was compared by Sobolewski et al. [5]. Rael Cahn used band powers for delta and gamma bands during Vipassana Meditation [6]. Asieh Ahani et al. analyzed MM using spectral and phase analysis for EEG and respiration signals [7]. Shao Wei Xue compared theta activity data at rest and after 1 week of Integrative Body–Mind Training (IBMT) using network analysis of EEG [8]. Fred Travis proposed automatic self-transcending meditation [9]. Asieh Ahani used EEG and respiration signals for MM and analyzed data using spectral analysis [10]. Decho Surangsrirat used a consumer EEG device for spectral analysis of data from Buddhist monks during meditation and other activities [11]. Antoine Lutz reviewed the Focused Attention (FA) and Open Monitoring (OM) meditation types [12]. Chamandeep Kaur reviewed different types of meditation techniques and signal processing challenges. Coherence parameter is useful to find the brain's functional connectivity [13]. Dissanayaka et al. examined the functional connectivity during awake, drowsiness and meditative state using three different methods: Directed Transfer Function (DTF), Welch, and Minimum Variance Distortion less Response (MVDR) [14]. Jadhav N. et al. explained the effect of meditation on emotional response using Hjorth parameters and discussed the effect of meditation on functional connectivity [15].

3 Materials and Methods

3.1 Subjects

11 healthy engineering volunteers in the age group of 21–23 were the subjects for this experiment. The consent form was signed by each of them before the experiment and subjects were informed about the experiment and meditation practice. Focused attention meditation practice for 20 min with spine erect was asked to do by the subjects for 8 weeks. In which, the focus on breathing, i.e., incoming and outgoing breath is to be done. In the experiment, the data was recorded by each subject for meditation of 5 min.

The protocol for data acquisition was approved by the ethics committee of Shri Guru Singhji Institute of Engineering and Technology, Nanded.

3.2 EEG Data Recording

To record the data, EMOTIV EPOC+ wireless EEG device with 14 electrodes was used. A separate setup was used for recording the EEG data. The sampling frequency of the device is 128 samples/s. The 14 electrodes of the device are based on the 10–20 system, which are AF3, AF4, F3, F4, FC5, FC6, F7, F8, T7, T8, P7, P8, O1, and O2 as shown in Fig. 1. The EEG data was recorded for 5 min meditation before and after 8 weeks of meditation practice.

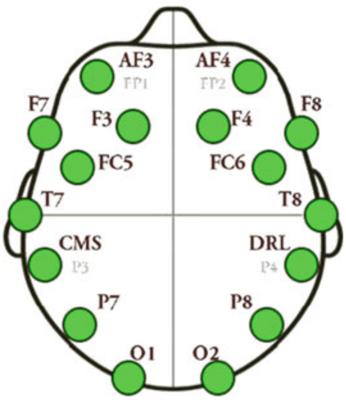


Fig. 1 EMOTIV headset

3.3 Preprocessing

Data with zero mean and one standard deviation was preprocessed in the range of 0.5–30 Hz frequency as shown in Eq. 1. The fourth-order IIR Butterworth band-pass filter was used to pre-process the data.

$$x_p = \frac{(x - m_x)}{\sqrt{\sigma_x}} \tag{1}$$

where m_x and $\sqrt{\sigma_x}$ are mean and standard deviation of input signal x.

3.4 Higher Order Crossings

All the time series possess a local as well as the global up and down movements with respect to time. A time series with zero mean which oscillates about zero level can be expressed using zero-crossing count. Applying a filter to a time series, changes the oscillations of time series which in turn changes the number of zero crossings. Using this concept, if we apply the filter to a given time series and count the number of zero crossings to time series iteratively, we will get a resulting zero-crossing count which can be referred as HOC [16, 17]. Calculation of HOC can be done by applying appropriate filter design. The procedure to calculate the HOC is given as:

1. Find the backward difference of a signal using a difference operator ∇ and is given as

$$\nabla X_t = X_t - X_{t-1} \tag{2}$$

2. The sequence of difference operator (high pass filter) can be defined as

$$\zeta_k = \nabla^{k-1} \tag{3}$$

3. The number of zero crossings can be estimated by constructing a binary time series $Z_t(k)$ as

$$Z_t(k) = \begin{cases} 1, & \zeta_k(X_t) \geq 0 \\ 0, & \zeta_k(X_t) < 0 \end{cases} \tag{4}$$

   where k = 1, 2, 3, …; t = 1, …, N.
4. The corresponding HOC is given as

$$D_k = NZC\{\zeta_k(X_t)\} \tag{5}$$

5. The value of D_k is calculated by counting the number of symbol changes in $Z_1(k)$, …, $Z_N(k)$,

$$D_k = \sum_{t=2}^{N}\left[Z_t(k) - Z_{t-1}(k)\right]^2 \tag{6}$$

The overall data of 5 min was segmented into the epochs of 5 s for analysis using HOC.

3.5 Functional Connectivity

The functional connectivity can be analyzed by using Magnitude-Squared Coherence Estimation (MSCE). The coherence is defined as the measure for the synchronization between brain regions. To calculate the coherence, Welch averaging method is used. The coherence is the ratio of cross spectral density of two variables to the product of auto spectral density of both variables. Magnitude-Squared Coherence is a function of frequency with values between 0 and 1 that indicates how well the electrode (i) corresponds to the electrode (j) of EEG at each frequency. The MSCE can be calculated as

$$M_{coh(ij)}(f) = \frac{\left|S_{ij}(f)\right|}{S_{ii}(f) \times S_{jj}(f)} \tag{7}$$

where $S_{ii}(f)$ and $S_{jj}(f)$ is auto spectral density, $|S_{ij}(f)|$ is cross power spectral density of electrode (i) and electrode (j) of EEG. The $S_{ii}(f)$ and $S_{jj}(f)$ and $|S_{ij}(f)|$ can be estimated as

$$|S_{ii}(f)| = |X_i(f)X_{i*}(f)| \tag{8}$$

$$\left|S_{jj}(f)\right| = \left|X_j(f)X_{j*}(f)\right| \tag{9}$$

$$\left|S_{ij}(f)\right| = \left|X_i(f)X_{j*}(f)\right| \tag{10}$$

where $X_i(f)$ and $X_j(f)$ are the Discrete Fourier Transform (DFT) of ith electrode signal ($X_i(t)$) and jth electrode signal ($X_j(t)$) respectively. The coherence of each electrode is calculated with all other electrodes [15].

The MSCE for each electrode is calculated with all other electrodes. The electrode pair having maximum MSCE is considered with its effective frequency.

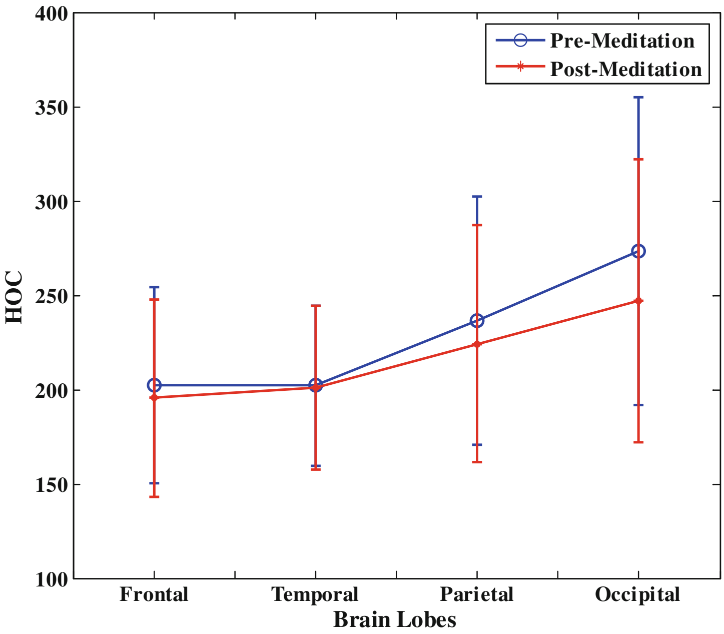


Fig. 2 HOC in pre- and post-meditation

4 Result and Discussion

The result is discussed for each feature individually.

4.1 Higher Order Crossings

The HOC (across order k = 1–9) for each lobe in pre- and post-experiment is shown in Fig. 2. The HOC in post-experiment is decreased for all lobes except temporal lobe. This result shows that the zero crossing in post-experiment is less, i.e., EEG signal variation is less. The HOC values for all brain lobes in pre- and post-mediation are compared using mean (±standard deviation) values. In premeditation, the values [frontal: 154.71 (±36.30), temporal: 156.47 (±28.52), parietal: 189.74 (±49.21), occipital: 235.93 (±66.76)] are more as compared to post-meditation [frontal: 150.48 (±36.97), temporal: 158.26 (±29.69), parietal: 175.63 (±45.52), occipital: 203.12 (±58.80)] (except temporal).

The frontal lobe is involved in emotion regulation, problem solving, speech, and movement. The decrease in HOC of frontal lobe indicates that meditation practice helps to regulate the emotions, improves the problem solving ability. The parietal lobe is also involved in problem solving, pain, and taste [2]. The decrease in HOC

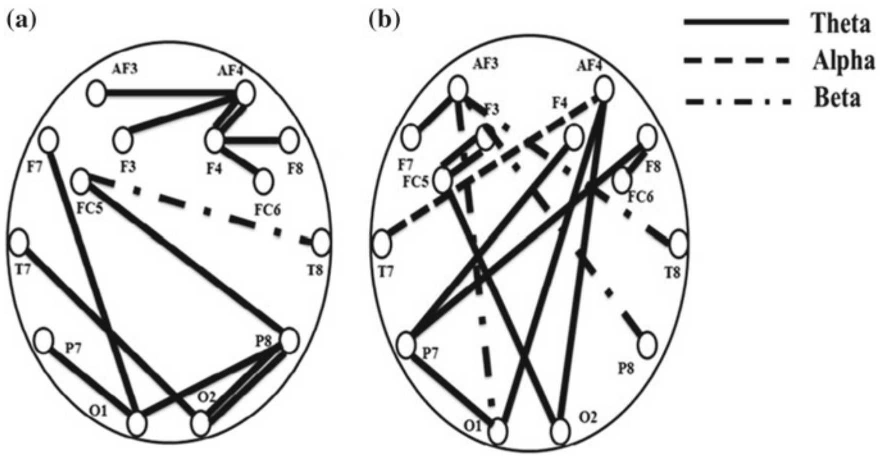


Fig. 3 Functional connectivity: **a** Pre-meditation, **b** Post-meditation

of this lobe indicates meditation may help to be calm in painful situation also. The role of temporal lobe is of hearing and memory. The increase in HOC of this lobe indicates the awareness of subject towards surrounding improves.

4.2 Functional Connectivity

The topographical representation of the functional connectivity is shown in Fig. 3. The functional connectivity is calculated for pre- and post-meditation state. The topographical representation shows difference in premeditation (Fig. 3a) and post-meditation (Fig. 3b) states, using this it can be considered that the functional connectivity can be used as a measure to verify the effect of meditation on brain. The effective frequency and effective electrode, given for each electrode is related to the maximum coherence with that particular electrode. Table 1 explains the effective electrode pair and the effective frequency in pre- and post-meditation, respectively.

As shown in Fig. 3, the dotted line represents the theta band (4–8 Hz), thick line represents the alpha band (8–12 Hz), and dash-dot line represents the beta band (12–30 Hz). In the post-meditation, the connectivity in between the right and left hemisphere (interhemispheric connectivity) is increased as compared to premeditation. Also from Table 1, it can be seen that while meditation the dominating frequency is in the range of alpha and beta band.

Table 1 Effective electrodes and frequency

Premeditation			Post-meditation		
Effective electrode (i)	Effective electrode (j)	Effective frequency (Hz)	Effective electrode (i)	Effective electrode (j)	Effective frequency (Hz)
AF3	AF4	10	AF3	O1	16
AF4	F4	9	AF4	O2	10
F7	O1	10	F7	AF3	11
F8	F4	9	F8	P7	10
F3	AF4	10	F3	FC5	11
F4	AF4	9	F4	P7	9
FC5	P8	10	FC5	F3	11
FC6	F4	9	FC6	F8	19
T7	O2	26	T7	AF4	5
T8	FC5	18	T8	AF3	20
P7	O1	9	P7	O1	9
P8	O2	10	P8	AF3	16
O1	P8	10	O1	AF4	10
O2	P8	10	O2	FC5	11

5 Conclusion

The reduction in HOC values of frontal and parietal lobe indicates reduction in mental load for problem solving and also regulation of emotion is possible. The increase in HOC value of temporal lobe indicates the subject became more aware of surrounding after meditation. The result of functional connectivity shows the increase in interhemispheric connectivity. So from the results, we can say that meditation helps to become calm and to hold attention on the desired purpose with minimal efforts.

Declaration The work reported in this chapter is approved by the ethical approval committee of SGGSIE&T, Nanded. The committee consist of Dr. Mrs. S. S. Shinde (Chairperson), Dr. S. T. Hamde, Dr. R. R. Manthalkar, Prof. A. K. Dhaolwe, Prof. A. K. Dhadve.

References

1. Davidson, R.J., Lutz, A.: Buddha's brain: neuroplasticity and meditation [in the spotlight]. IEEE Signal Process. Mag. **25**(1), 176–174 (2008)
2. Nidal, K., Malik, A.S.: EEG/ERP Analysis: Methods and Applications. Crc Press (2014)
3. Lo, P.-C., Zhu, Q.: Microstate analysis of alpha-event brain topography during Chan meditation. In: 2009 International Conference on Machine Learning and Cybernetics. IEEE (2009)
4. Chang, K.-M., Lo, P.-C.: F-VEP and alpha-suppressed EEG-physiological evidence of inner-light perception during Zen meditation. Biomed. Eng. Appl. Basis Commun. **18**(01), 1–7 (2006)

5. Sobolewski, A., et al.: Impact of meditation on emotional processing—a visual ERP study. Neurosci. Res. **71**(1), 44–48 (2011)
6. Cahn, B.R., Polich, J.: Meditation states and traits: EEG, ERP, and neuroimaging studies. Psychol. Bull. **132**(2), 180 (2006)
7. Ahani, A., et al.: Quantitative change of EEG and respiration signals during mindfulness meditation. J. Neuroeng. Rehabil. **11**(1), 87 (2014)
8. Xue, S.-W., et al.: Short-term meditation induces changes in brain resting EEG theta networks. Brain Cogn. **87**, 1–6 (2014)
9. Travis, F., Shear, J.: Focused attention, open monitoring and automatic self-transcending: categories to organize meditations from Vedic, Buddhist and Chinese traditions. Conscious. Cogn. **19**(4), 1110–1118 (2010)
10. Ahani, A., et al.: Change in physiological signals during mindfulness meditation. In: 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER). IEEE (2013)
11. Surangsrirat, D., Intarapanich, A.: Analysis of the meditation brainwave from consumer EEG device. In: SoutheastCon 2015. IEEE (2015)
12. Lutz, A., et al.: Attention regulation and monitoring in meditation. Trends Cogn. Sci. **12**(4), 163–169 (2008)
13. Kaur, C., Singh, P.: EEG derived neuronal dynamics during meditation: progress and challenges. Adv. Prev. Med. **2015** (2015)
14. Dissanayaka, C., et al.: Comparison between human awake, meditation and drowsiness EEG activities based on directed transfer function and MVDR coherence methods. Med. Biol. Eng. Comput. **53**(7), 599–607 (2015)
15. Jadhav, N., Manthalkar, R., Joshi, Y.: Effect of meditation on emotional response: an EEG-based study. Biomed. Signal Process. Control **34**, 101–113 (2017)
16. Petrantonakis, P.C., Hadjileontiadis, L.J.: Emotion recognition from EEG using higher order crossings. IEEE Trans. Inf. Technol. Biomed. **14**(2), 186–197 (2010)
17. Kedem, B., Yakowitz, S.: Time Series Analysis by Higher Order Crossings. IEEE press, New York (1994)

The Detrended Fluctuation Analysis of EEG Signals: A Meditation-Based Study

Sunil R. Hirekhan, Ramchandra Manthalkar and Shruti Phutke

Abstract The Detrended Fluctuation Analysis is a widely used method for analysis of non-stationary time series which has been applied to EEG signals. The Detrended Fluctuation Analysis (DFA) of the EEG signals in pre- and post-meditation (mindfulness) intervention are compared. It is observed that the EEG data obtained from 8 subjects out of total 11 subjects shows reduction in the DFA values. The reduction in DFA values represents the lower intrinsic fluctuations in the EEG time series, which is a measure of better (higher) complexity of these vital rhythms. The reduced DFA values after 8 weeks of Focused Attention (mindfulness) meditation practice in more number of subjects, indicates that the meditation practice enhances the ability to handle complexity. The reduced DFA values indicate improved neuronal functioning of these subjects.

Keywords Detrended fluctuation analysis (DFA) · Power-law correlation
Mindfulness

1 Introduction

The long-range correlation in the EEG data is an important parameter, and can act as a complexity indicator of the neuronal system. In order to distinguish long-range intrinsic fluctuations in the data, it is essential to discard any trends existing in the data which otherwise may lead to detection of false long-range correlations in time series. A direct calculation of the correlation may not be appropriate in case of any noise superimposed on the data, or in presence of trends in the data. The Detrended Fluctuation Analysis (DFA) is considered to be an established method for determining the scaling behaviour of noisy data in presence of trends [1]. It is a

S. R. Hirekhan (✉) · R. Manthalkar · S. Phutke
Department of Electronics and Telecommunication, SGGSIE&T, Nanded, India
e-mail: hirekhansunil@yahoo.com

R. Manthalkar
e-mail: rrmanthalkar@sggs.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_78

convenient method to analyse the non-stationary signals and quantifies fractal like scaling properties of EEG signal. It is a modified Root Mean Square method, which computes the root mean square distance of the signal from the local trend line as a function of scale parameter [2].

In this work, the DFA method has been used as it has been successfully applied for the analysis of many of the human physiological processes [3–5]. The DFA method is a powerful technique used by the researchers to distinguish a healthy human system when compared to a diseased one, and is applied as a complexity metric [6]. The traditional theory suggests that the human physiological systems achieve homeostasis, reflected in terms of the complexity of the physiological signals. A healthy system automatically maintains (a rise) complexity of the signals, while a diseased one is not able to keep it. The rise in the complexity is indicated by the reduction in alpha (α) values of the log–log plot of the $F_{DFA}(s)$ versus (s) of the physiological time series [5, 6].

1.1 Background

It may be noted that the Detrended Fluctuation Analysis (DFA) method invented by Peng et al. [3] was utilised for analysing the ECG signals of healthy and CVD patients, and observed a higher mean value of scaling exponent ($\alpha = 1.24$) for subjects with CVD, compared to the ($\alpha = 1.00$) healthy subjects [3]. Mathieu Jospin et al. have applied the DFA technique to measure the depth of anaesthesia injected to patients [4]. Zebende et al. have applied DFA technique to analyse EEG signals and have found that the rms DFA function is greater for the frontal channels than the parietal channels [7]. Richard Hardstone et al. proposed that the EEG system is scale-free dynamics, by applying DFA methods for the analysis of the same [8]. Manuel Varela et al. have applied the DFA method to measure the complexity of glucose time series, and examined the characteristics of crossover point in it to investigate the risk of developing type-2 diabetes in subjects [5]. Amir Weissman et al. by applying DFA techniques determined whether the fluctuations in glucose blood levels pose fractal behaviour [9]. Carmen Rodríguez de Castro et al. have found DFA method useful in analysing the glucose profile, and stated that the DFA method have significantly performed as a harbinger of type-2 diabetes development in high risk population [6]. Daniel Ab´asolo et al. have studied the EEG background activity of Alzheimer's disease (AD) patients to find that the scaling behaviour of EEG is sensitive to AD [10].

Previously, a number of approaches are adopted by the researchers for the analysis of vital physiological signals, but recently, detrended fluctuation analysis has been popularly used for the same. In this work, for the first time, application of the DFA method for the analysis of Meditation-based EEG time series is implemented. It is observed that the results obtained are in confirmation with the conclusions derived by earlier researchers [3, 4].

2 Materials and Methods

The EEG data have been recorded for 11 healthy subjects (6 males and 5 females), age group 22–29 (± 3.19 yrs) who participated voluntarily in the pre- and post-meditation experiment. Though, the pre-intervention data was collected for 32 participants, most of these subjects could not maintain regular meditation practice. The data samples from these subjects were discarded from post-intervention analysis, and only 11 subjects who have maintained the regular meditation practice, are considered for data analysis. EEG data is recorded using EMOTIVE EPOCH+, 14-channel EEG recording device. The Focused Attention (mindfulness) meditation was practiced by the subjects daily for 20 min for a period of 8-weeks. The experimental protocol was approved by the ethical committee of this institute, SGGSIE&T, Nanded.

2.1 Detrended Fluctuation Analysis: Brief Description

The EEG time series is denoted as {x(t)}, t: No. of time points (ranging from 0 to N; N = 37888)

The following steps are performed:

1. Average value of time series, $\overline{x(t)}$, is subtracted from x(t), as follows.

$$\mathrm{y(n)} = \sum_{t=1}^{n} \left[\{x(t)\} - \overline{x(t)}\right\}, \mathrm{n} = 1, \ldots, \mathrm{N}; \tag{1}$$

2. Divide the entire time range in p equal windows, discarding any remainder, such that each window has s = int(N/p) time points, window size.
(Each window size varies from 74 to 18944 time points, with No. of boxes = 512 to 02, respectively. The maximum window size may be half length of input data, with two adjacent intervals [11]).

3. Within each window, P (P = 1, …, p), perform a least square fit of {y(n)} by a straight line $y_P(n)$ to determine the local trend of the Pth window.
(This step is implemented using the “detrend” operation of the MATLAB 2014b, which removes best straight-line fit from the input vector, in the given window, maintaining intrinsic fluctuations in the original EEG time series).

4. The Variance of the intrinsic fluctuation, {y(n)} from the trend line $y_P(n)$ in *P*th window is, $F^2(s)$

$$\mathrm{F}^2(\mathrm{s}) = \frac{1}{s} \sum_{n=(P-1)s+1}^{Ps} [\{y(n)\} - y_P(n)]^2; \tag{2}$$

It is measure of local detrended fluctuation in P^{th} window.

5. The square root of the average fluctuation of $F^2(s)$ overall windows is the RMS fluctuation from the local trend in p windows, each of s time points (DFA),

$$F_{DFA}(s) = \sqrt{\frac{1}{p}\sum\nolimits_{P=1}^{p} F^2(s)}; \tag{3}$$

The above computations for a single EEG electrode are performed for 14 electrodes and repeated for 11 subjects, pre- and post-interventions. The algorithmic complexity of the DFA technique is of linear complexity, in terms of big O notation, is O(n). As the number of EEG electrodes increases, either or number of subjects increases, the time taken for computation of DFA shall increase proportionately.

The study of dependence of $F_{DFA}(s)$ on window size, s, is the essence of DFA.

If there is a power-law behaviour, $F_{DFA}(s) \propto s^{\alpha}$, where α is an indicator of the nature of fluctuations in EEG time series [10].

If $\alpha < 0.5$, the correlations in the signal are anti-persistent, i.e. an increment is very likely to be followed by a decrement, and vice versa;

If $\alpha = 0.5$, time series represents uncorrelated white noise;

If $\alpha > 0.5$, the correlations in the signal are persistent, i.e. an increment is very likely to be followed by an increment and vice versa.

The exponent, α, estimated by DFA lies between 0 and 2.

The detrending operation maintains only the intrinsic fluctuations in the EEG data with the removal of trends. The detrended EEG signal for the EEG electrode AF4 for a window size of 592 samples, and no. of boxes = 64, is shown in Fig. 1.

It may be observed that the EEG time series of all the 14 electrodes acquired in this experiment follows a power-law relationship. The intrinsic fluctuation of the EEG data varies with the size of the window. The exponent, α of this correlation, which is the slope of the linear graph of the log–log plot, log ($F_{DFA}(s)$) versus log(s) lies (almost) between 0.5 and 1.0. As EEG time series follows the power-law behaviour, the intrinsic fluctuation in the data increases with the size of the window, as seen from Fig. 2.

It may be observed that the fluctuations in the EEG data is lower in magnitude (Y-axis) when window size (s) is small (sample size, s = 74) and the fluctuations in

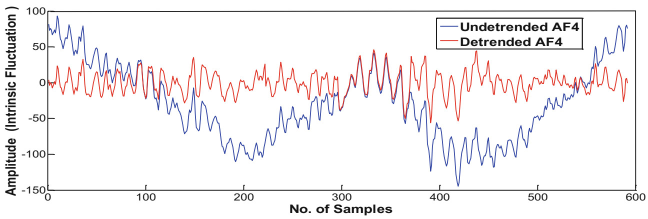


Fig. 1 Frontal electrode (AF4): undetrended and detrended signal

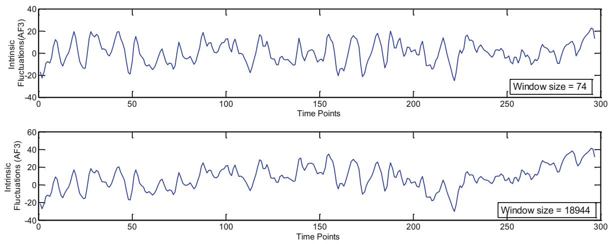


Fig. 2 Frontal electrode (AF3) signal in a window size of 74 time points and in a window size of 18944 time points

the EEG data are higher in magnitude when window size (s) is large (sample size, s = 18944), for the same EEG data segment (No. of time points = 296) on the X-axis. Similar is the nature of behaviour of the EEG data, which can be readily observed in case of all EEG electrode signals.

A complex time series will have a greater amount of information in the small details, and therefore their description will become rapidly more and more inaccurate as the graining (time-window size) increases [6]. Similarly, less complex time series will retain a more accurate description in coarser graining (large size windows). DFA, the intrinsic fluctuation is the measure of the difference between the curve (time series) and the regression line (trend). The area $F_{DFA}(s)$ can be interpreted as the "gap" between the territory (time series) and the "map" (linear regression). Complex time series will quickly increase their "map versus territory gap" as the graining (time-window size) increases. Consequently, complex series will have a less steep log ($F_{DFA}(s)$) versus log(s) slope, and a lower DFA [6]. This has been observed in human physiological systems, such as gluco-regulation, thermoregulation, neurologic disorders, heart rhythms, etc.

Basically, DFA explores the complexity of a time series, analysing the rate of information loss as the graining of the description become coarser (larger). Hence, though all the EEG data series may exhibit the power-law nature, but the rate at which the intrinsic fluctuation changes with the window size indicates the complexity of the EEG data series.

In this work, the Detrended Fluctuation Analysis method is considered for the analysis of Meditation-based EEG signals. The Focused Attention (mindfulness) Meditation technique in which the focus of attention need to be maintained on the incoming and outgoing breath, has been practised for 8 weeks, daily for 20 min. The EEG data of all 11 subjects who participated voluntarily in the Meditation experiment have been analysed pre and post-intervention. The intrinsic fluctuations of EEG signals of all the electrodes, i.e. the DFA values are computed for all the subjects.

The power-law correlation is given as

$$F_{DFA}(s) \propto s^{\alpha} \tag{4}$$

Table 1 DFA values: before and after meditation (Subject-1 –Subject-5)

Elect rode	SB1	SA1	SB2	SA 2	SB3	SA3	SB4	SA4	SB5	SA5
AF3	23.48	23.94	11.3	6.76	3.25	4.55	3.17	4.33	8.83	10.00
AF4	34.36	15.54	6.90	11.06	6.54	6.20	3.68	8.91	9.75	7.98
F3	26.52	50.90	8.41	5.75	2.47	4.10	3.09	3.90	9.16	9.62
F4	37.87	15.82	8.35	4.97	17.91	3.14	2.32	3.66	9.67	11.39
F7	30.99	10.47	4.06	5.28	2.34	3.22	2.13	3.26	8.87	10.73
F8	21.38	19.06	9.52	5.66	2.49	3.94	2.63	3.62	9.43	3.78
FC5	35.67	27.60	7.52	4.89	2.59	7.55	3.77	7.93	9.07	8.98
FC6	28.65	12.17	6.89	5.54	3.53	4.33	4.75	7.56	8.77	8.90
O1	56.41	11.95	4.80	6.58	2.80	6.17	3.32	4.21	10.5	7.99
O2	51.92	17.47	5.17	7.67	18.20	3.49	4.12	4.38	9.84	8.33
P7	33.10	16.94	10.8	5.84	4.81	13.96	4.56	4.48	10.3	9.01
P8	25.37	54.08	4.61	5.60	2.95	3.48	3.82	4.48	10.0	9.41
T7	25.62	20.87	8.66	14.18	6.91	6.73	4.32	4.81	13.2	9.36
T8	54.78	45.61	8.86	7.83	4.62	4.39	3.81	4.67	12.6	8.93
Avg.	**34.72**	**24.46**	**7.57**	**6.97**	**5.82**	**5.38**	3.53	5.01	10.0	8.89

Table 2 DFA values: before and after meditation (Subject-6–Subject-11)

Elec-trode	SB6	SA6	SB7	SA7	SB8	SA8	SB9	SA9	SB10	SA1 0	SB11	SA1 1
AF3	12.95	4.3	24.2	4.9	9.8	10.2	11.5	4.26	7.04	5.5	5.46	16.9
AF4	5.46	4.1	16.5	7.5	9.6	9.40	11.3	3.82	40.9	6.1	7.01	15.2
F3	2.66	3.0	52.2	7.1	7.7	8.99	12.0	3.56	27.0	3.7	5.23	15.3
F4	2.50	2.8	17.3	5.3	6.6	7.47	10.9	3.68	8.27	8.0	6.18	12.2
F7	3.77	5.2	13.7	3.4	4.1	8.11	10.9	2.53	14.0	4.7	10.1	16.1
F8	2.68	1.8	29.8	4.2	4.4	5.55	8.24	1.95	4.79	6.2	3.30	1.74
FC5	2.32	2.4	10.1	9.3	5.3	8.91	9.55	2.98	11.0	5.6	3.99	9.83
FC6	2.20	3.1	13.6	5.4	6.4	18.9	9.19	5.17	10.1	5.8	5.21	11.5
O1	9.28	2.6	21.9	5.1	4.7	5.62	16.8	97.8	5.63	5.5	5.06	12.7
O2	6.01	2.6	16.4	5.8	6.6	8.96	9.55	3.85	15.6	5.9	7.65	15.8
P7	7.14	3.6	19.0	5.2	6.7	9.01	11.8	3.49	8.68	5.6	5.73	18.9
P8	1.74	4.2	24.6	5.9	6.4	9.09	11.0	3.39	16.3	4.5	5.94	13.3
T7	6.28	4.8	19.5	12.9	10.7	9.42	12.1	4.50	11.4	8.52	8.98	16.5
T8	18.9	3.9	19.3	4.9	8.8	10.5	10.7	3.9	22.8	5.81	7.26	15.8
Avg	5.99	3.4	**21.3**	**6.2**	7.0	9.30	**11.1**	**10.3**	**14.5**	**5.84**	6.22	13.7

$$\mathrm{Log}(F_{DFA}(s)) = \alpha\,\mathrm{Log}(s) \tag{5}$$

It may observed that the high complexity EEG series changes at a small rate, hence the DFA values obtained shall be small, and the slope of the linear relationship in Eq. (1), i.e. α is small. In this regard, the following table depicts the DFA values of the subjects after the meditation intervention.

As shown in Tables 1 and 2, column SB (before meditation) and column SA (after meditation) are the DFA values for subject-1 to subject-5 and subject-6 to subject-11, respectively. It can be observed that the average DFA values decreases for 8 subjects amongst total 11 subjects, which are shown in "blue" colour.

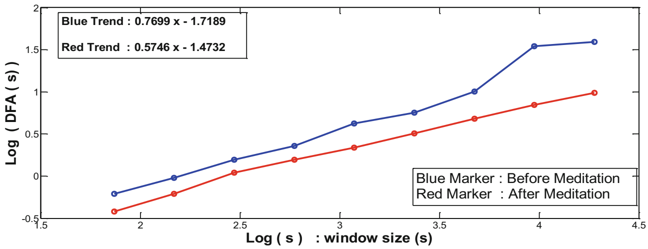


Fig. 3 Intrinsic fluctuation versus window size (log–log) plot

3 Results and Discussion

The intrinsic fluctuation for the parietal electrode (P8), before and after meditation, is shown in Fig. 3. The exponent "α" which is the slope of the linear graph log–log plot of the $F_{DFA}(s)$ versus window size (s) is reduced after the meditation intervention as represented by equations in text box. For log ($F_{DFA}(s)$) plot shown in red colour, the slope (α2) of this linear graph is 0.5746 (after meditation), whereas for log ($F_{DFA}(s)$) plot shown in blue colour, the slope of the linear graph (α1) is 0.7699. This indicates that the rate of change of intrinsic fluctuations (α) of EEG time series, and hence the DFA values are reduced after Focused Attention meditation practice.

This change in slope of the log–log plot of the $F_{DFA}(s)$ versus window size (s) for all 14 EEG electrodes, for all 11 subjects is computed. The electrode-wise change in "α" exponent at various electrode positions is observed, as depicted in following tables.

It may be observed in Table 3 for subject-1 to subject-6 and in Table (4) for subject-7 to subject-11, column SB (before meditation) and column SA (after meditation) depicts in "blue" colour, the EEG electrode positions at which the reduction in DFA exponent, alpha (α) has taken place after Meditation intervention. It may be noted that for the subject-5 and subject-9, the exponent alpha (α) is reduced for all the 14 electrodes. That is, the subjects must have strongly maintained the meditation practice. The number of EEG electrodes at which reduction in exponent alpha (α) is observed is as given in Table 5 below.

It can be observed that out of 11 subjects, total 9 subjects have achieved reduction in alpha (α) exponent in more than 50% (i.e. 7 EEG electrodes), or higher number of electrodes after meditation intervention. This indicates that the efficiency of the Autonomic Nervous System (ANS), which maintains the balance between para-sympathetic and sympathetic activities are enhanced as reflected by rise in the complexity of the brain rhythms in more number of subjects through the practice of Focused Attention meditation.

The DFA analysis of EEG data supports the potential benefits of meditation practices. In the future, similar analysis shall be carried out for higher number of subjects, in presence of a control group. Further, data from the long-term meditators shall be analysed for further investigation.

Table 3 DFA exponent (α): before and after meditation

Electrode	Subject-1		Subject-2		Subject-3		Subject-4		Subject-5		Subject-6	
	SB1	SA1	SB2	SA2	SB3	SA3	SB4	SA4	SB5	SA5	SB6	SA6
AF3	0.9	0.91	0.70	**0.61**	0.55	**0.49**	0.53	**0.50**	0.72	**0.61**	1.02	**0.61**
AF4	0.96	**0.80**	0.50	0.75	0.66	**0.56**	0.55	**0.51**	0.71	**0.62**	0.71	**0.65**
F3	0.88	0.92	0.66	**0.60**	0.64	**0.57**	0.51	**0.48**	0.74	**0.64**	0.69	**0.63**
F4	0.97	**0.79**	0.79	**0.69**	1.01	**0.57**	0.52	**0.51**	0.76	**0.65**	0.65	0.69
F7	0.87	**0.76**	0.64	0.75	0.59	0.63	0.55	0.58	0.84	**0.67**	0.76	**0.52**
F8	0.84	0.86	0.83	**0.77**	0.63	**0.61**	0.48	0.58	0.62	**0.55**	0.61	0.68
FC5	0.99	**0.88**	0.57	0.62	0.63	0.85	0.46	0.52	0.81	**0.66**	0.62	0.62
FC6	0.96	**0.81**	0.52	0.56	0.63	0.69	0.42	0.50	0.78	**0.63**	0.54	0.57
O1	0.97	**0.78**	0.58	0.73	0.58	0.66	0.45	0.50	0.83	**0.61**	0.75	**0.58**
O2	0.95	**0.86**	0.59	0.65	0.87	**0.57**	0.51	0.53	0.77	**0.63**	0.75	**0.56**
P7	0.91	**0.69**	0.78	**0.67**	0.60	0.86	0.61	**0.57**	0.77	**0.68**	0.85	**0.60**
P8	0.89	**0.85**	0.61	0.63	0.58	**0.54**	0.54	**0.50**	0.73	**0.63**	0.61	0.73
T7	0.87	0.91	0.76	**0.75**	0.65	**0.52**	0.54	**0.50**	0.77	**0.64**	0.70	**0.60**
T8	1.04	1.05	0.83	**0.62**	0.64	**0.47**	0.60	**0.51**	0.77	**0.60**	1.06	**0.62**

Table 4 DFA exponent (α): before and after meditation

Electrode	Subject-7		Subject-8		Subject-9		Subject-10		Subject-11	
	SB7	SA7	SB8	SA8	S B9	SA9	SB10	SA10	SB11	SA11
AF3	0.831	**0.687**	0.619	**0.572**	0.747	**0.504**	0.733	**0.705**	0.603	0.711
AF4	0.893	**0.689**	0.734	**0.668**	0.753	**0.570**	1.071	**0.629**	0.645	0.757
F3	1.164	**0.736**	0.624	**0.624**	0.736	**0.556**	1.029	**0.686**	0.599	0.698
F4	0.970	**0.644**	0.648	0.659	0.785	**0.622**	0.894	**0.578**	0.689	0.843
F7	1.024	**0.677**	0.655	0.709	0.814	**0.562**	0.943	**0.597**	0.883	**0.792**
F8	1.021	**0.602**	0.641	0.764	0.811	**0.627**	0.741	**0.480**	0.659	**0.650**
FC5	0.740	0.793	0.646	0.758	0.787	**0.604**	0.774	**0.675**	0.586	**0.546**
FC6	0.854	**0.591**	0.589	0.844	0.759	**0.632**	0.654	0.701	0.593	0.687
O1	1.01	**0.655**	0.589	0.735	0.813	**0.712**	0.653	0.655	0.617	0.712
O2	0.964	**0.685**	0.575	0.691	0.790	**0.574**	0.896	**0.581**	0.660	0.743
P7	0.929	**0.749**	0.647	**0.639**	0.901	**0.690**	0.805	**0.610**	0.638	0.849
P8	0.529	0.623	0.635	**0.619**	0.769	**0.574**	0.952	**0.620**	0.624	0.734
T7	0.888	**0.825**	0.678	**0.642**	0.770	**0.642**	0.819	**0.621**	0.741	**0.725**
T8	0.942	**0.658**	0.565	0.617	0.764	**0.519**	0.962	**0.651**	0.625	0.740

4 Conclusion

The regular practice of Focused Attention meditation for a span of 8 weeks have resulted in the reduced DFA exponent alpha (α), and reduced DFA values. The reduction in DFA values and alpha (α) exponent after the meditation intervention shows reduction in the rate of change of intrinsic fluctuation, within same window size. This indicates an increase in the complexity of the EEG series, which is an indicator of enhancement in mental functioning of the brain. The above analysis has

Table 5 Number of EEG electrodes which had reduction in DFA exponent alpha (α)

Number of electrodes	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7	Sub8	Sub9	Sub10	Sub11
	9	7	9	8	14	9	12	6	14	12	4

been supported by the computation of Hurst exponent using rescaled range analysis, which indicates potential benefits of meditation practices.

Declaration The work reported in this chapter is approved by the ethical approval committee of SGGSIE&T, Nanded. The committee consists of Dr. Mrs. S. S. Shinde (Chairperson), Dr. S. T. Hamde, Dr. R. R. Manthalkar and Prof. A. K. Dhadve.

References

1. Kantelhardt, J.W., Koscielny-Bunde, E., Rego, H.H.A., Havlin, S., Bunde, A.: Detrended long-range correlations with detrended fluctuation analysis. Physica A: Stat. Mech. Appl. **296**, 441–454 (2001)
2. Adda, A., Benoudnine, H.: Detrended fluctuation analysis of EEG recordings for epileptic seizure detection. In: IEEE International Conference on Bio-engineering for Smart Technologies (BioSMART), pp. 1–4 (2016)
3. Peng, C.K., Havlin, S., Stanley, H.E., Goldberger, A.: Quantification of scaling exponents and crossover phenomena in non-stationary heartbeat time series. CHAOS **5**(1), 85–87 (1995)
4. Jospin, M., Caminal, P., Jensen, E.W., Litvan, H., Vallverdú, M., Struys, M.M.R.F., Vereecke, H.E.M., Kaplan, D.T.: Detrended fluctuation analysis of EEG as a measure of depth of anesthesia. IEEE Trans. Biomed. Eng. **54**(5), 840–846 (2007)
5. Varela, M., Vigil, L., Rodriguez, C., Vargas, B., García-Carretero, R.: Delay in the detrended fluctuation analysis crossover point as a risk factor for type 2 diabetes mellitus. J. Diabetes Res. 1–6 (2016)
6. Rodríguez de Castro, C., Vigil, L., Vargas, B., Delgado, E.G., Carretero, R.G., Ruiz-Galiana, J., Varela, M.: Glucose time series complexity as a predictor of type 2 diabetes. Diabetes/Metab. Res. Rev. **33**, 1–9 (2017)
7. Zebende, G.F., Oliveira Filho, F.M., Leyva Cruz, J.A.: Auto-correlation in the motor/imaginary human EEG signals: a vision about the *FDFA* fluctuations. PLoS ONE **12**(9), 1–13 (2017)
8. Hardstone, R., Poil, S.S., Schiavone, G., Jansen, R., Nikulin, V.V., Mansvelder, H.D., Linkenkaer-Hansen, K.: Detrended fluctuation analysis: a scale-free view on neuronal oscillations. Front. Physiol. **3**(450), 1–13 (2012)
9. Weissman, A., Binah, O.: The fractal nature of blood glucose fluctuations. J. Diabetes Complicat. **28** 646–651(2014)
10. Abasolo´, D., Hornero, R., Escudero, J., Espino, P.: A study on the possible usefulness of detrended fluctuation analysis of the electroencephalogram background activity in Alzheimer's disease. IEEE Trans. Biomed. Eng. **55** (9) 2171–2179 (2008)
11. Phothisonothai, M., Nakagawa, M.: Fractal-based EEG data analysis of body parts movement imagery tasks. J. Physiol. Sci. **57**(4) 217–226 (2007)

Convex Optimization-Based Filter Bank Design for Contact Lens Detection

Swati Madhe and Raghunath Holambe

Abstract We have designed a novel convex optimization-based filter bank (FB), which minimizes the frequency band errors and optimizes time–frequency localization at the same time. The designed FB is regular and satisfies the constraint of perfect reconstruction (PR). In convex optimization, we have optimized quadratic constrained quadratic programs by transforming it into a semidefinite program. We have also compared the frequency band errors and time–frequency localization of proposed FB with existing FB. We have used this FB for designing a new contact lens detection (CLD) system. The IIITD database has been used for this purpose. The results have been expressed in terms of correct classification rate (CCR). The superiority of the designed FB has been shown by comparing the results with other existing CLD systems. The newly designed FB can also be effectively used for various signal processing applications.

Keywords Filter bank · Convex optimization · Frequency band errors

1 Introduction

Multiresolution analysis using filter banks (FBs) and wavelets has become an integral part of many signal processing applications such as data compression, watermarking, etc. FBs with linear phase filters are essential in many applications. However, linear phase filters are present only in the case of Haar FBs along with orthogonality. Hence, biorthogonal FBs are selected over orthogonal FBs in these applications. Perfect reconstruction filter banks (PRFBs) are generated from wavelet bases. Wavelets can be constructed by iterating PRFBs, if the wavelet bases satisfy the

S. Madhe (✉)
Cummins College of Engineering, Pune 411052, India
e-mail: swatimadhe@gmail.com

R. Holambe
SGGS College of Engineering and Technology, Nanded 431606, India
e-mail: rsholambe@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_79

additional constraint of regularity. FBs are developed using different design techniques. Biorthogonal FBs have linear phase characteristics and hence, are preferred in image processing applications over orthogonal FBs. PR and symmetry can be achieved in biorthogonal FB easily. The designing of FB keeping all the constraints in mind is a challenging task. Since not much research has been done in this field, we have endeavored to venture into this area and come up with a unique FB design which satisfies most of the criteria.

The PRFBs can be designed with different methods such as spectral factorization [1, 2], polyphase and lattice schemes [3, 4], and parameterization [5]. These strategies have additionally been utilized to optimize desired FB properties. In this paper, we have designed a biorthogonal FB using convex optimization considering the frequency band errors of FB as an objective function and time–frequency localization and PR as constraints. Convex optimization gives a global solution and also allows us to incorporate various constraints. The designed FB can be used in any signal processing application. In this paper, we have demonstrated the use of the designed FB for Contact Lens Detection (CLD).

2 Literature Review

FBs are designed by considering various properties of filters such as energy compaction, flatness, regularity, orthogonality, linear phase, and frequency band errors depending on applications [6]. The block diagram of a 1dimensional (1D) two-channel FB is shown in Fig. 1. This FB is characterized by its analysis ($H_0(z)$ and $H_1(z)$) and synthesis ($G_0(z)$ and $G_1(z)$) filters. $H_0(z)$ and $G_0(z)$ aim to be ideal low pass filters (LPFs) with a cut off frequency of 90° and $H_1(z)$ and $G_1(z)$ aim to be ideal high-pass filters (HPFs) with a cutoff frequency of 90°. In the past, two-channel FB based on time–frequency localization was designed mainly using polynomial factorization, lifting scheme, and optimization. In case of polynomial factorization [1], frequency responses of filters are not explicitly controlled. Ansari et al. [7] used lifting scheme to construct a class of a triplet half-band filter bank (THFB). It has better frequency selectivity but these FBs are not regular. FBs designed using optimization techniques are numerically efficient and gives a global solution [8]. FB design based on optimization techniques involves formulating objective function using the property of FB. We have considered frequency band errors and the time–frequency localization in our FB design.

Iris is a standout among the most encouraging biometric modalities, and is regularly used in high-security applications because of its high accuracy. Daughman's approach was used in all commercial iris recognition systems and is still the frequently used approach. Recent studies have proved that although iris recognition is the most accurate biometrics, the accuracy is reduced due to several covariates such as pupil dilation and use of different cameras for capturing iris images. In addition to the abovementioned covariates, the presence of contact lenses is also a major factor in reducing the overall accuracy and effectiveness of the iris recognition system.

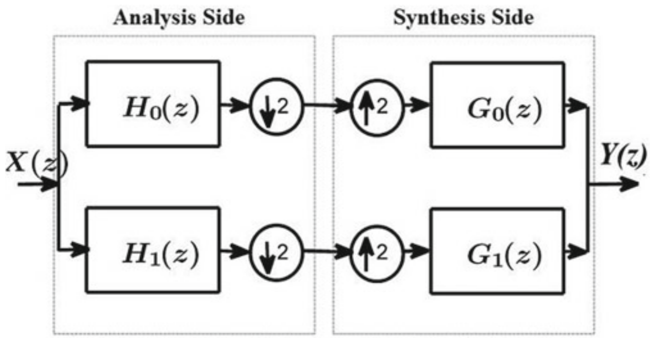


Fig. 1 Block diagram of one dimensional two-channel filter bank

Thus, to improve the efficiency of the identification system, it becomes imperative to detect contact lenses [9]. Hence, we have worked towards designing a CLD system which can not only detect contact lens but also classify it as No lens (N), Soft lens (S), and Texture lens (T). This CLD system has been shown to be more accurate than the existing ones.

The inequality of variables can be easily solved with convex optimization problem as the objective function. Constraint functions can also be convex. Convex optimization is preferred over local optimization methods in many applications because of the following advantages [10]:

- It ensures that infeasible solutions are not generated in searching for an optimum.
- All local optima are global optima.
- Algorithms written to solve convex optimization problems are faster, more efficient and very reliable.

Convex optimization is obtained using linear programming. There has been a recent improvement in computing and in optimization theory. Semidefinite relaxation (SDR) technique is used in many optimization problems for reformulation as convex minimization problems. In non-convex quadratically constrained quadratic programs (QCQP) optimization can be achieved effectively using SDR techniques as described in [11]

$$\begin{gathered}\min_{\boldsymbol{x}} \boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{B}\boldsymbol{x}\\ \text{subject to}\, \boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{C}_j\boldsymbol{x} \geq g_j,\ j = 1, 2, \ldots, p\\ \boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{D}_j\boldsymbol{x} \geq l_j,\ j = 1, 2, \ldots, q\end{gathered} \tag{1}$$

where $\boldsymbol{B}, \boldsymbol{C}_1, \ldots, \boldsymbol{C}_p, \boldsymbol{D}_1, \ldots, \boldsymbol{D}_q$ are symmetric matrices, possibly indefinite, the vector $\boldsymbol{x}$ contains optimization variables and $g_1, \ldots, g_p, l_1, \ldots, l_q$ are constants [12].

In [13–15], the authors have designed FBs with optimum time–frequency localization value. Nonlinearity and non-convexity are involved in the optimization problem of designing optimally time–frequency localized FB with minimum frequency band errors. The design problem considered here is optimization of QCQP by transforming it into SDP. Recently, Sharma et al. [16] have used convex optimization to design optimally time–frequency localized FB. The authors have minimized the frequency variance of the filter for the predefined time variance or minimized the time variance keeping the frequency variance fixed subjected to the constraint of PR. The proposed approach optimizes the time–frequency localization of the filters. But the authors have not considered the frequency band errors in the design. Reducing frequency band errors is important in FB design as it assures that the designed filter responses are similar to the ideal filter responses. Hence in our design, we have used convex optimization for designing FBs with the objective function of minimizing the frequency errors with the predefined frequency and time variances used as equality constraints with PR.

3 Design Methodology

In this section, we present the SDR technique used in convex optimization to design the FB. The objective function is to minimize the frequency band errors of the analysis LPF and synthesis LPF considering the constraint of PR and time–frequency localization.

3.1 Formulation of Objective Function

Consider a linear phase LPF with impulse response h(n) and order 2M. Figure 2 shows desired response D(w) and the designed response H(w). The filter frequency response is expressed as

$$\mathrm{H(w)} = \mathrm{h(0)} + \sum_{\mathrm{n}=1}^{\mathrm{M}} 2 \cdot \mathrm{h(n)} \cdot \cos(\mathrm{nw}) = \mathrm{x}^{\mathrm{T}}\mathrm{c(w)} \tag{2}$$

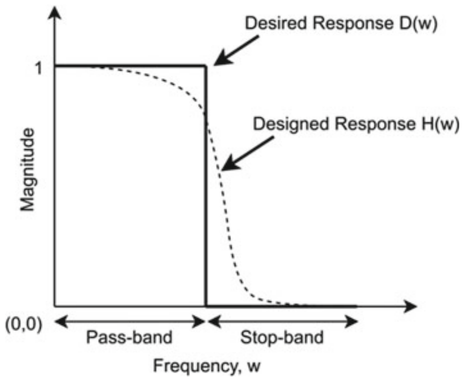
where, vectors x, c(w) $\in \mathrm{R}^{\mathrm{M+1}}$ are defined as

$$\mathrm{x} = [\mathrm{h(0)h(1)h(2)} \ \ldots \ \mathrm{h(M-1)\,h(M)}]^{\mathrm{T}}, \text{ and}$$

$$\mathrm{c(w)} = [1 \ \cos(\mathrm{w})2\cos(2\mathrm{w}) \ \ldots \ 2\cos(\mathrm{Mw})]^{\mathrm{T}} \tag{3}$$

The desired response D(w) given by

Fig. 2 Frequency response of desired and designed LPF



$$\begin{aligned} \mathrm{D(w)} &= 1, \quad 0 \le \mathrm{w} \le \mathrm{w}_1 \\ &= 0, \quad \mathrm{w}_2 \le \mathrm{w} \le \pi \end{aligned} \tag{4}$$

where w_1 and w_2 are e pass-band and stop-band cutoff frequencies, respectively. The stop-band error can be defined as

$$E_s = \int_{w_2}^{\pi} [D(w) - H(w)]^2 dw \tag{5}$$

$$E_s = \boldsymbol{x}^{\boldsymbol{T}}. \int_{ws}^{\pi} [c(w)] \cdot [c(w)]^T dw \cdot \boldsymbol{x} \tag{6}$$

The pass-band error can be expressed as

$$E_p = \boldsymbol{x}^{\boldsymbol{T}} \cdot \int_{0}^{w_1} [c(0) - c(w)] \cdot [c(0) - c(w)]^T dw \cdot \boldsymbol{x}$$

The objective function is written as the total of the pass-band and stop-band errors

$$\varphi = \beta E_p + (1 - \beta) E_s, \quad \beta \in [0, 1]$$

$$\varphi = \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{E} \boldsymbol{x} \tag{7}$$

where β is the trade-off factor and $\boldsymbol{E}$ is the total error.

3.2 Formulation of Constraints

3.2.1 Perfect Reconstruction (PR) Constraint

Perfect reconstruction means $Y(z)$ is exactly the replica of $X(z)$. For the PR constraint formulation, we express product filter $R(z)$ as,

$$R(z) = H_0(z)G_0(z) \tag{8}$$

The design of two-channel PRFB reduces to the design of the half-band filter $R(z)$. For PRFB synthesis, LPF $g_0(n)$ should be the complement of analysis LPF $h_0(n)$. The coefficients of $g_0(n)$ are given by $g_0 = [g_0(0), g_0(1), \ldots, g_0(N)]$, where $2N$ is the order of the filter. Hence, the PR constraint in linear form is

$$Rg_0 = 0 \tag{9}$$

where analysis LPF $g_0(n)$ is used to form the real matrix R.

3.2.2 Time–Frequency Localization Constraint

Considering a linear phase LPF with impulse response $f(n)$ and order $2P$. The time and frequency variances of $f(n)$ in $l_2(z)$ normalized to unit energy are defined in [17] as

$$\sigma_t^2 = \sum n^2 |f(n)|^2 = \frac{1}{\pi}\int_0^{\pi} \left|F'(w)\right|^2 dw = \boldsymbol{x^T T x}$$

$$\sigma_f^2 = \frac{1}{\pi}\int_0^{\pi} w^2 |F(w)|^2 dw = \boldsymbol{x^T F x} \tag{10}$$

where **T** and **F** are real, positive-definite, symmetric matrix of size (P + 1) × (P + 1) formed using the filter response equations. (We have referred [15] for details.)

The objective function of minimizing frequency band errors is subjected to the constraints of fixed time–frequency localization and PR (9).

3.3 Design of Analysis LPF

Considering the objective function of frequency band errors and the constraints, the design of analysis LPF is expressed as QCQPs

$$\min_{\boldsymbol{H}} \boldsymbol{\varphi} = Tr(\boldsymbol{EH}),\ \text{Minimize frequency band errors}$$

$$\begin{aligned}\text{Subject to}\quad & Tr(\boldsymbol{TH}) = C_1,\ \text{Fixed time variance constraint}\\ & Tr(\boldsymbol{FH}) = C_2,\ \text{Fixed frequency variance constraint}\end{aligned}$$

where Tr is a trace of a matrix and $\boldsymbol{H}$ is a new variable defined as $\boldsymbol{H} = h_0 h_0^T$.

3.4 Design of Synthesis LPF

Using the PR constraint, the design of synthesis LPF is expressed as QCQPs

$$\min_{\boldsymbol{H}} \boldsymbol{\varphi} = Tr(\boldsymbol{EH}),\ \text{Minimize frequency band errors}$$

$$\begin{aligned}\text{Subject to}\quad & Tr(\boldsymbol{TG}) = C_3,\ \text{Fixed time variance constraint}\\ & Tr(\boldsymbol{FG}) = C_4,\ \text{Fixed frequency variance constraint}\\ & Tr(\boldsymbol{RG}) = 0,\ \text{PR constraint}\end{aligned}$$

where G is a new variable defined as $\mathrm{H} = \mathrm{g}_0 \mathrm{g}_0^{\mathrm{T}}$. The SDPs are solved using convex optimization toolbox CVX [18]. The installation and implementation details are given in [19].

4 Design Example and Comparison of the Properties

The designed FB is illustrated with a design example. Design parameters are $M = 3$, $N = 6$, $\beta = 0.5$. The time variance and frequency variance of the analysis filter are taken as 0.7 and 0.4, respectively. The time variance of the synthesis filter to be designed is fixed to 0.4 and frequency variance is less than 1.5. The responses of the designed filters are shown in Fig. 3.

The proposed design is compared with the existing method given in [16]. The comparison in terms of time–frequency localization and frequency band errors between the proposed method and existing FB is shown in Table 1. It is observed that the frequency band errors are minimized and the time–frequency localization is optimized as compared to the existing FB. It is because we have considered the frequency band errors as the objective function and considered the time and frequency variance as the constraints along with PR.

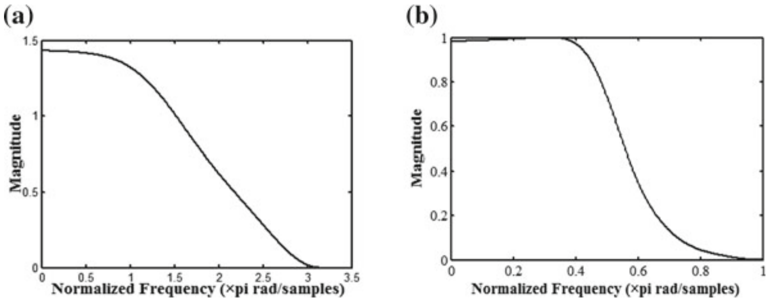


Fig. 3 Design example frequency response. **a** Analysis LPF. **b** Synthesis LPF

Table 1 Comparison of designed FB with existing FB

Filter	Property	FB [15]	Designed FB
Analysis LPF	Time–frequency localization	0.3268	**0.3061**
	Frequency band error	64.759	**51.542**
Synthesis LPF	Time—frequency localization	0.8005	**0.51542**
	Frequency band error	20.306	**18.2682**

5 Contact Lens Detection (CLD) System Using Designed FB

The accuracy of any iris recognition system is negatively affected due to the presence of contact lenses [20]. The designed FB can effectively extract all the features necessary for detecting contact lenses. Support Vector Machine (SVM) classifier have been used to classify lenses into three classes No lens (N), Soft lens (S), and Texture lens (T). The performance of the designed system has been evaluated by comparing its correct classification rate (CCR) with the CCR of existing system.

Table 2 and Table 3 show the CCR of IIIT-D database compared with LBP and mLBP algorithm in [21] for intra-camera and multi-camera case experimentations, respectively. In case of intra-camera, the datasets have been tested separately. On the other hand, for multi-camera, the datasets have been combined and tested. The results show that the CCR of the CLD system designed using the new FB is more accurate than the existing techniques. The improved results can be attributed to the newly designed FB.

Table 2 CCR of intra-camera case experimentation using proposed FB

Dataset	Class	LBP + SVM [21]	mLBP [21]	Proposed FB
IIITD Cogent	N	65.53	66.8	**97.5**
	S	42.73	56.6	**96.6**
	T	89.39	94.9	**95.8**
	Average	65.88	73.0	**96.8**
IIITD Vista	N	53.37	76.2	**97.6**
	S	50.9	67.5	**99.2**
	T	98.64	91.6	**99.8**
	Average	67.63	80.0	**99.5**

Table 3 CCR of multi-camera case experimentation using proposed FB

Dataset	Class	mLBP [21]	Proposed FB
IIITD Database	N	62.14	**74.47**
	S	61.63	**65.82**
	T	94.74	**95.24**
	Average	72.84	**78.51**

6 Conclusion

In this paper, a new approach to design an FB with minimized frequency band errors and optimum time–frequency localization has been introduced using convex optimization. The PR constraint has also been considered in the design. This FB has better time–frequency localization and reduced frequency band errors. It can be easily converted into 2D by using a tensor product. Hence, the designed FB can be effectively used to process images in various applications such as feature extraction, image fusion, watermarking, etc.

References

1. Patil, B.D., Patwardhan, P.G., Gadre, V.M.: On the design of fir wavelet filter banks using factorization of a halfband polynomial. IEEE Signal Process. Lett. **15**, 485–488 (2008)
2. Rahulkar, A.D., Patil, B.D., Holambe, R.S.: A new approach to the design of biorthogonal triplet half-band filter banks using generalized half-band polynomials. Signal Image Video Process. **8**(8), 1451–1457 (2014)
3. Vaidyanathan, P.P., Hoang, P.-Q.: Lattice structures for optimal design and robust implementation of two-channel perfect-reconstruction QMF banks. IEEE Trans. Acoust. Speech Signal Process. **36**(1), 81–94 (1988)
4. Vaidyanathan, P.P.: Multirate digital filters, filter banks, polyphase networks, and applications: a tutorial. Proc. IEEE **78**(1), 56–93 (1990)
5. Phoong, S.-M., Kim, C.W., Vaidyanathan, P.P., Ansari, R.: A new class of two-channel biorthogonal filter banks and wavelet bases. IEEE Trans. Signal Process. **43**(3), 649–665 (1995)

6. Daubechies, I.: Ten lectures on wavelets. SIAM (1992)
7. Ansari, R., Kim, C.W., Dedovic, M.: Structure and design of two-channel filter banks derived from a triplet of halfband filters. IEEE Trans. Circuits Syst. II: Analog Digit. Signal Process. **46**(12), 1487–1496 (1999)
8. Patil, B.D., Patwardhan, P.G., Gadre, V.M.: Eigenfilter approach to the design of one-dimensional and multidimensional two channel linear-phase fir perfect reconstruction filter banks. IEEE Trans. Circuits Syst. I: Regul. Pap. **55**(11), 3542–3551 (2008)
9. Daugman, J.G.: High confidence visual recognition of persons by a test of statistical independence. IEEE Trans. Pattern Anal. Mach. Intell. **15**(11), 1148–1161 (1993)
10. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
11. Goemans, M.X., Williamson, D.P.: Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. J. ACM (JACM) **42**(6), 1115–1145 (1995)
12. Luo, Z.-Q., Ma, W.-K., So, A.M.-C., Ye, Y., Zhang, S.: Semidefinite relaxation of quadratic optimization problems. IEEE Signal Process. Mag. **27**(3), 20–34 (2010)
13. Morris, J.M., Peravali, R.: Optimum duration discrete-time wavelets. Opt. Eng. **36**(4), 1241–1248 (1997)
14. Sharma, M., Kolte, R., Patwardhan, P., Gadre, V.: Time-frequency localization optimized biorthogonal wavelets. In: 2010 International Conference on Signal Processing and Communications (SPCOM), pp. 1–5. IEEE (2010)
15. Sharma, M., Gadre, V.M., Porwal S.: An eigenfilter-based approach to the design of time-frequency localization optimized two-channel linear phase biorthogonal filter banks. Circuits Syst. Signal Process. **34**(3) (2014)
16. Sharma, M., Bhati, D., Pillai, S., Pachori, R.S., Gadre, V.M.: Design of time–frequency localized filter banks: transforming non-convex problem into convex via semidefinite relaxation technique. Circuits, Syst. Signal Process. **35**(10), 3716–3733 (2016)
17. Tay, D.B.: Balanced-uncertainty optimized wavelet filters with prescribed regularity. In: Proceedings of the 1999 IEEE International Symposium on Circuits and Systems, ISCAS'99, vol. 3, pp. 532–535 (1999)
18. Grant, M., Boyd, S., Ye, Y.: CVX: Matlab software for disciplined convex programming (2008)
19. Grant, M., Boyd, S., Ye, Y.: CVX users' guide (2009)
20. Kohli, N., Yadav, D., Vatsa, M., Singh, R.: Revisiting iris recognition with color cosmetic contact lenses. In: 2013 International Conference on Biometrics (ICB), pp. 1–7. IEEE (2013)
21. Yadav, D., Kohli, N., Doyle, J.S., Singh, R., Vatsa, M., Bowyer, K.W.: Unraveling the effect of textured contact lenses on iris recognition. IEEE Trans. Inf. Forensics Secur. **9**(5), 851–862 (2014)

EEG Waveform Classification Using Transform Domain Features and SVM

Hemprasad Y. Patil, Priyanka B. Patil, Seema R. Baji and Rohini S. Darade

Abstract Electroencephalogram (EEG) waveforms are fluctuations in brain-recorded utilizing anodes set on the scalp. Albeit a few strategies for the evaluation of working of brain, for example, MEG, PET, CT scan, and MRI have been presented, the EEG waveform is as yet an important biological signal for checking the brain signal variations because of its moderately ease and being helpful for the patient. We have presented an approach to classify the EEG waveforms into two classes, viz. epileptic and normal. The algorithm fuses the features extracted using discrete wavelet transform, discrete cosine transform, and stationary wavelet transform. The fused features are subjected to support vector machine (SVM) classifier.

Keywords EEG signals · DWT · DCT · SVM · SWT · Machine learning

1 Introduction

Electroencephalography (EEG) is a popular neuroimaging method for logging the activities inside human brain that are ordinarily utilized to study the subtleties of biological neural networks, and identify brain ailments. Since a huge EEG waveforms

H. Y. Patil (✉)
R. H. Sapat College of Engineering, Management Studies and Research,
Nashik, Maharashtra, India
e-mail: hemprasadpatil@gmail.com

P. B. Patil
Sir Dr. M. S. Gosavi Polytechnic Institute, Nashik Road, Nashik, Maharashtra, India
e-mail: patilpriyanka1991@gmail.com

S. R. Baji · R. S. Darade
Late G. N. Sapkal College of Engineering, Anjaneri, Trimbakeshwar Road,
Nashik, Maharashtra, India
e-mail: bajiseema1@gmail.com

R. S. Darade
e-mail: rohini.darade16@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_80

are being logged and for a human being, it is difficult to examine the EEG waveforms by observation. In this view, it becomes necessary to employ computers with machine learning algorithms to extract meaningful information which will lead to decisions. Major steps in EEG processing toolchain comprises of waveform preprocessing (filtering), feature extraction to yield feature vectors, and classification using machine learning techniques [1].

Epilepsy is a precarious neurological ailment which occurs due to impermanent anomalous emancipations of the electrical currents which flow in the brain, which results in involuntary fluctuations and tremors in the various parts of human body [2]. Epilepsy has a position of second topmost occurring neurological ailment in the world [2]. Although it has lot many patients all over the world, epilepsy happens to be a perplexing task to identify and medicate. The people with epilepsy are more prone to subsequent neurological diseases, hence it becomes crucial to identify such patients and treat them properly [2].

As a neuron's activation function is processed, it hits with neighboring neuron. Due to this progression, a trifling aggregate of electric discharge is created [3]. The resultant waveform has small amplitude which makes it very problematic to calculate its frequency. Therefore, numerous electrodes are positioned on the human brain's scalp to record the neural electrical movement [3]. The logged waveforms are time-varying, and not stationary. This demands a very careful signal processing and machine learning techniques to come to a conclusion about class of the particular waveform. As an initial step, the recording probes are positioned on diverse areas of human scalp, namely occipital area, frontal area, temporal area, and parietal area [3].

This paper is organized as follows. Section 1 contains introduction to EEG signal recording and need of epilepsy ailment classification. Section 2 presents a literature survey about the existing EEG classification techniques as well as methods and materials such as DWT, DCT, and SWT. Section 3 describes proposed method for EEG classification. Section 4 presents results and discussion. The conclusion is presented in Sect. 5.

2 Related Work and Background

Amin et al. [1] have utilized the intrinsic DWT based temporal EEG feature extraction techniques for classification of the EEG waveforms. In this approach, the following steps are followed: (i) EEG signals were registered through Raven's advance progressive metric test and (ii) the EEG signals reported in position of eyes open of the patient. They have achieved resultant accuracy of above 98% using the SVM, MLP, and the K-NNC.

Ma et al. [4] have presented a technique which utilizes the nature-inspired PSO algorithm for optimization of the penalty metrics and kernel choice for achieving the better results from support vector machines. The computed results authenticate that assumption of using the optimized SVM classifier for segregation of motor imagery based EEG signals.

Subasi [5] demonstrated the use of artificial neural network for supervised learning in the context of EEG signal classification. The ME ANN network has been trained to get the two discrete class outcomes, namely normal person and epileptic person. Gating functions are been employed to yield more recognition accuracy. The ME network architecture has recorded better accuracy rates than a single stand-alone neural network model.

Guo et al. [6] have used the nature-inspired genetic programming algorithms to obtain features from database of EEG waveforms for enhancing the biased accuracy of a utilized classifier and reduction in input feature dimensionality at the same time. The branch assembly of genetic programming gives the required features, and a new computed function gives the selective features.

Jahankhani et al. [7] described the usage of artificial neural network models for EEG wave classification. The computed neural model was being examined for various classification accuracies and the obtained results portray that the proposed algorithm has been successful in EEG signal classification.

2.1 Discrete Cosine Transform (DCT)

As soon as natural signals proceed to visual information processing areas of the brain, they are mapped as contrast (absolute part), phase part, and frequency information which are aspects of transform domain analysis [8]. Due to data compression which is a major characteristic of the discrete cosine transform, we are motivated to use it as a part of feature vector to represent EEG waveform.

The mathematical equations for DCT are as follows [9]. For the input discrete signal $x[n]$ comprising of N samples, its corresponding DCT $dct(k)$ is given in (1–2).

$$dct(k) = \alpha(k) \sum_{n=0}^{N-1} x[n] \cos \frac{(2n+1)\pi k}{2N} \tag{1}$$

For $0 \leq k \leq N-1$,
where

$$\alpha(0) = \sqrt{\frac{1}{N}}, \quad \text{and} \quad \alpha(k) = \sqrt{\frac{2}{N}} \quad \text{For,} \ 1 \leq k \leq N-1 \tag{2}$$

2.2 Discrete Wavelet Transform (DWT)

The notion behind wavelet transform is the multi-resolution (MR) decomposition of real-life signals. The common methodology for MR analysis is to generate an approximation part of a signal with wavelet scaling function (typically LPF) and detailed parts with wavelet mappings (typically HPF).

The signal processing operations such as translation and dilation on mother wavelet yields the family of basis function mappings [10]. The mathematical inner product of original waveform and generated wavelet functions gives DWT coefficients. The DWT coefficients at single decomposition stage can be calculated as in (3–4).

$$DWT_L(n, j) = \sum_m DWT_L(m, j-1)h(m-2n) \tag{3}$$

$$DWT_H(n, j) = \sum_m DWT_L(m, j-1)g(m-2n) \tag{4}$$

2.3 Stationary Wavelet Transform (SWT)

For the calculation of stationary wavelet transform (SWT) at each decomposition level, after application of LPF and HPF to the original signal, decimation is not performed. This results in the creation of wavelet coefficients that are of same length as original signal. For obtaining this transformation, scaling function (LPF) as given in (5) is used [11].

$$\phi_{m,n}(x) = 2^{-m}\phi(2^{-m}x - n) \tag{5}$$

If the wavelet function denoted by $\varphi(x)$, the wavelet coefficients [11] after decomposition are given by (6).

$$w_{m,n} = \left\langle f(x), 2^{-m}\phi(2^{-m}x - n)\right\rangle \tag{6}$$

3 Proposed Method

The EEG logged database utilized for this experimentation is a subset dataset, which is developed by Dr. Andrzejak at the University of Bonn [12]. This database is logged using a 128-channel ADC and amplifier prototype system which is comprised of 100 single-channel EEG waveforms with a sampling frequency of 173.61 Hz and for 23.6 s duration per subset. For this experiment, we have considered the subsets of normal patients and seizure activity. The sample feature realization from normal set and epileptic set are displayed in Fig. 1 and Fig. 2, respectively. The feature variations across the samples of normal and epileptic signals are captured in these figures.

The block diagram of the proposed system is shown in Fig. 3.

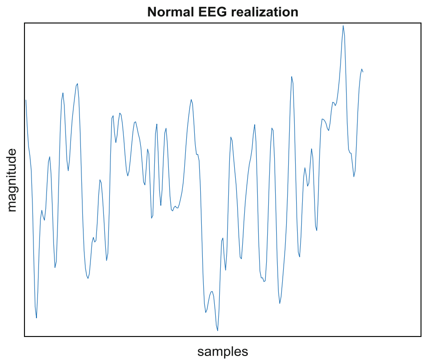


Fig. 1 Sample feature realization from normal person

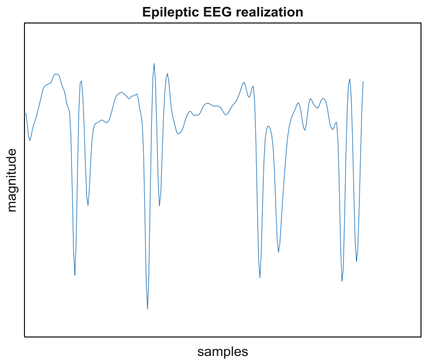


Fig. 2 Sample feature realization from epileptic person

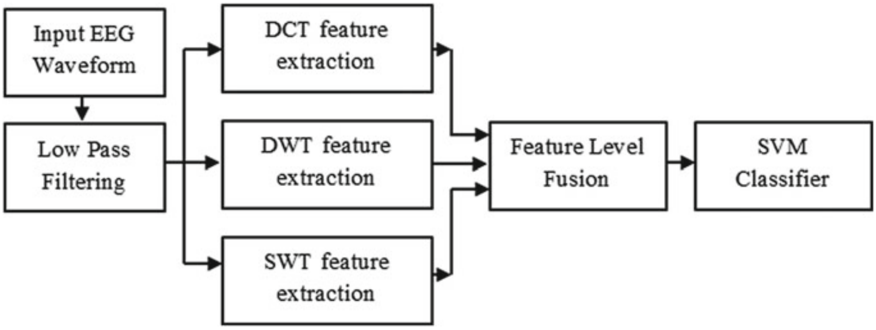


Fig. 3 Proposed system for EEG signal classification

The input EEG signal from subsets of database is filtered with FIR low pass filter. Subsequently, it is passed to DCT, DWT, and SWT blocks for transform domain realization. The support vector machine (SVM) is utilized for classification purpose. The algorithm of the proposed method is given below.

Step 1: Import raw EEG data (normal and epileptic) in MATLAB.
Step 2: With the help FFT, apply low pass filtering to remove irrelevant artifacts.
Step 3: Split up signals with window of length 256.
Step 4: Construct feature vectors using fusion of DWT, SWT, and DCT-based features. The DWT and SWT features consist of mean, maximum, minimum, standard deviation, and metrics. Additional skewness features from DWT contribute to accuracy.
Step 5: Split the data in separate partition which will be used for training and testing and is mutually exclusive.
Step 6: Perform classification using SVM classifier.
Step 7: Measure performance using confusion matrix statistics such as accuracy, sensitivity and specificity.

4 Results and Discussion

The step-wise experimentation indicates the relevance of feature level fusion for better performance. The performance measures are given by (7–9). The results are illustrated in Table 1.

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

The performance measures with this experiment are summarized in Fig. 4.

Table 1 Performance measures

Sr. No.	Experimental case	Accuracy (%)	Sensitivity (%)	Specificity (%)
1	Discrete Wavelet Transform (DWT)	99.50	99.50	99.50
2	Stationary Wavelet Transform (SWT)	99.69	99.62	99.62
3	DWT and SWT	99.81	99.62	99.62
4	DWT with Skewness and SWT	99.81	99.62	99.62
5	DWT with Skewness, SWT and DCT	99.88	99.75	99.75

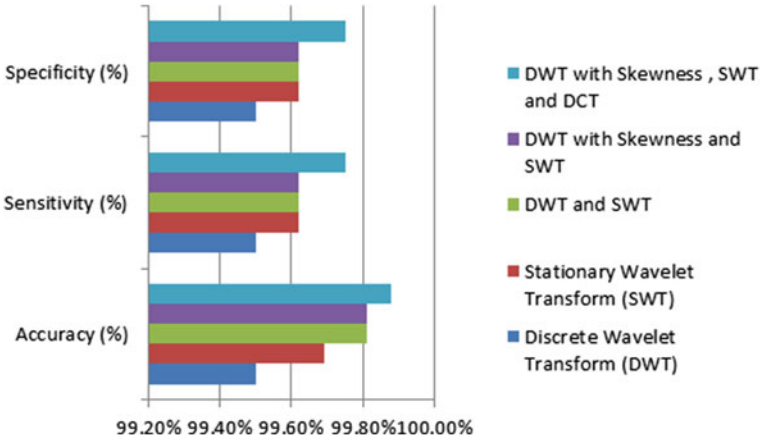


Fig. 4 Performance measures used in experimentation

Table 2 Comparison with reported methods

Sr. No.	Reference	Training: Testing Samples	Specificity (%)	Sensitivity (%)	Accuracy (%)
1	Subasi [5] MLPNN	1000:600	92.60	93.60	93.20
2	Subasi [5] ME	1000:600	94.00	95.00	94.50
3	Proposed	800:800	99.75	99.75	99.88

The EEG signal classification using Transform domain methods has been performed in a step-wise manner to study the effect of DWT, SWT, and DCT on the performance metrics. It has been observed that the fusion of DWT with skewness, SWT, and DCT features yields higher performance as compared with individual metrics. Furthermore, the benchmarking with EEG classification by Subasi [5] has been done as indicated in Table 2, which indicates that, with less number of samples in training, better performance can be obtained with feature level fusion of DWT with skewness, SWT, and DCT. Most of the EEG signals are nonstationary in nature. Transform domain techniques are found to be suitable for analysis of such nonstationary signals. This was the motivation behind selection of transform domain methods and cost functions for EEG feature extraction.

5 Conclusion

EEG signal classification is an important challenge which may act as a decision support system to the health practitioners. As epilepsy is widespread ailment, this proposed system may be widely useful. We have proposed a fusion of DCT, DWT and SWT based features which are subsequently subjected to SVM classifier. We have achieved accuracy of 99.88%, Sensitivity of 99.75%, and Specificity of 99.75%.

References

1. Amin, H., Malik, A., Ahmad, R., Badruddin, N., Kamel, N., Hussain, M., Chooi, W.: Feature extraction and classification for EEG signals using wavelet transform and machine learning techniques. Australas. Phys. Eng. Sci. Med. **38**(1), 139–149 (2015)
2. Hamad, A., Houssein, E., Hassanien, A., Fahmy, A.: Feature extraction of epilepsy EEG using discrete wavelet transform. In: 12th International Computer Engineering Conference (ICENCO), pp. 190–195. IEEE, Cairo (2016)
3. Satapathy, S., Dehuri, S., Jagadev, A.: ABC optimized RBF network for classification of EEG signal for epileptic seizure identification. Egypt. Inform. J. **18**(1), 55–66 (2017)
4. Ma, Y., Ding, X., She, Q., Luo, Z., Potter T., Zhang, Y.: Classification of motor imagery EEG signals with support vector machines and particle swarm optimization. Comput. Math. Methods Med. (2016)
5. Subasi, A.: EEG signal classification using wavelet feature extraction and a mixture of expert model. Expert Syst. Appl. **32**(4), 1084–1093 (2007)
6. Guo, L., Rivero, D., Dorado, J., Munteanu, C., Pazos, A.: Automatic feature extraction using genetic programming: an application to epileptic EEG classification. Expert Syst. Appl. **38**(8), 10425–10436 (2011)
7. Jahankhani, P., Kodogiannis, V., Revett, K.: The EEG symposium on modern computing signal and classification using the wavelet feature extraction and neural networks. In: John Vincent Atanasoff 2006 International Symposium on Modern Computing (JVA'06), pp. 120–124. IEEE, Bulgaria (2006)
8. Hafed, Z., Levine, M.: Face recognition using the discrete cosine transform. Int. J. Comput. Vis. **43**(3), 167–188 (2001)
9. Ahmed, N., Natarajan, T., Rao, K.: Discrete cosine transform. IEEE Trans. Comput. **100**(1), 90–93 (1994)
10. Grzeszczak, A., Mandal, M., Panchanathan, S.: VLSI implementation of discrete wavelet transform. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **4**(4), 421–433 (1996)
11. Wang, X., Istepanian, R., Song, Y.: Microarray image enhancement by denoising using stationary wavelet transform. IEEE Trans. Nanobiosci. **2**(4), 184–189 (2003)
12. Andrzejak, R., Lehnertz, K., Mormann, F., Rieke, C., David, P., Elger, C.: Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. Phys. Rev. E **64**(6), 0619071–0619078 (2001)

Colour-Adaptive Digital Image Watermarking Technique

Shailesh Sapkal and B. G. Hogade

Abstract Copyright protection and owner authentication have become necessary due to the circulation of large number of documents, images, audios, and videos through the internet. Manipulations and duplications in multimedia files and documents are very easy due to advancement in the signal and image processing algorithms. Therefore, it is very much important to devise watermarking techniques that are robust against geometrical distortions and collision attacks. In this paper, we proposed an adaptive digital image watermarking technique through colour features and Arnold transform in wavelet domain. An attempt has been made to adaptively transform the watermark image into a set of textures that visually match the colour of the input host image since colour features are invariant with respect to translation and rotation of the image. We first separate the three colours R, G, and B of the host and watermark image. Next, we decomposed the separated images using wavelet transform into sub-bands. Finally, low-frequency sub-band of the watermark image is embedded into low-frequency sub-band of the host image using Arnold transform. Experimental results on multiple host images and under various attacks using PSNR and correlation coefficient, clearly demonstrates that the proposed algorithm is robust and can be applied in colour image watermarking.

Keywords Adaptive watermarking technique · Colour feature · Arnold transform · Collision attacks

1 Introduction

Rapid growth in high-speed computer networks and internet of things (IoT) has explored means of new business, scientific, entertainment, and social opportunities. Electronic publishing and advertising, information delivery, data sharing, messaging, networking among computers, online shopping, banking, digital repositories and

S. Sapkal (✉) · B. G. Hogade
Department of Electronics Engineering, Terna College of Engineering, Navi Mumbai, India
e-mail: shails016@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_81

libraries, newspapers and magazines, online video, and audio, personal communication, and lots more have gained due to advancement in the internet. Transmission of images, audios, and videos over the internet is greatly enhanced due to cost effectiveness and improvement in technology. Computers have simplified storage and manipulations of digital data that includes text, images, audios, and videos with very high quality. This has resulted in a copy of a digital media which is exactly similar to the original. With similar copies of digital multimedia distribution over the internet, authentications and ownership are at more threat than ever due to the possibility of unlimited copying [1, 2].

Watermarking has been proved to be an efficient tool for the protection of Intellectual Property Rights (IPR) of image, audio, and video contents. Multimedia documents can be reproduced, changed, transformed, and diffused very easily through sophisticated signal and image processing algorithms. In this context, it is important to develop a system for copyright protection, protection against copying, and validation of contents. Digital watermark is a text/image/signature which is embedded into copyright digital content such as audios, videos, images, or text. Digital watermark can be detected or extracted later on if required. Such text/image/signature mostly carry the copyright or ownership identity of the content [3–5]. The process of embedding digital watermark information into digital content is known as digital watermarking.

Currently, the research in digital watermarking has been focused on mixing spatial with transformed domains. Transformed domain approaches such as DFT, DWT, and DCT concepts and adding more and more mathematical and statistical model. Also, other interdisciplinary approaches in watermarking: for example, use of chaotic theory, fractal image coding, and adaptive are explored. The wavelet transform decomposes the image into three spatial directions, i.e., horizontal, vertical, and diagonal. Wavelet transform is computationally efficient and can be implemented by using simple filter convolution [6–8]. It is generally observed that magnitude of DWT coefficients is larger in the lowest bands (LL) at each level of decomposition as compared to other bands (HH, LH, and HL). Thus, larger the magnitude of the wavelet coefficient the more significant it becomes for watermarking applications [9–12]. It is also observed that the extraction of watermark at lower resolutions is computationally effective since very few frequencies are involved at every successive resolution level. Edges, textures, and patterns in an image are easily located through high-resolution sub-bands. Arnold transform is commonly employed in two-dimensional domains for scattering of the pixels in an image. It destroys the relationship between the pixels which are evenly distributed, provides additional security to watermark, and also increases robustness of watermarking algorithm [13]. Principle component analysis (PCA) along with wavelet coefficients is applied mostly to blind or semi-blind watermarking techniques to acquire some features for decoding purpose [14]. PCA technique has resulted in very robust feature extraction and data extraction.

In this paper, we proposed adaptive digital image watermarking technique through colour features and Arnold transform in wavelet domain. An attempt has been made to adaptively transform the watermark image into a set of textures that visually match the colour of the input host image since colour features are invariant with respect to translation and rotation of the image. We first separate the three colours R, G, and B of the host and watermark image. Next, we decomposed the separated images using wavelet transform into sub-bands. Finally, low-frequency sub-band of the watermark image is embedded into low-frequency sub-band of the host image using Arnold transform. The paper is organized as follows Sect. 1 introduces to watermarking, Sect. 2 describes embedding and extraction algorithm in details. Results are discussed in Sect. 3 and finally concluded in Sect. 4.

2 Watermark Embedding and Extraction Algorithm

In this section, watermark embedding algorithm through colour features using Arnold transform and wavelet transform is discussed along with extraction algorithm. Watermark extraction is mostly the reverse process of embedding.

2.1 Embedding Algorithm

Let *I and w* represent the 8-bit colour input host and watermark image of the size $m \times n$. The steps involved in the embedding of the watermark image into host image are as follows:

1. Read the input host and watermark image of the size $m \times n$.
2. Separate the R, G, and B colour channels for both the host and watermark images.
3. Each colour channel R, G, and B of the host image was decomposed into four sub-bands LL, LH, HL and HH using Daubechies two-dimensional discrete wavelet transform.
4. Similarly each colour channel R, G, and B of the watermark image was decomposed into four sub-bands LL, LH, HL, and HH using Debauches two-dimensional discrete wavelet transform.
5. Determine the number of iterations i_{max} and initialized counter $i = 0$ for Arnold transform.
6. LL sub-band of the watermark image for all channels (R, G, and B) was scrambled using Arnold transform through equation

$$\left[m', n'\right] = [1, 1; 1, 2]\,[m, n] \mod 2 \tag{1}$$

7. Increment the counter $i = i + 1$.
8. Repeat steps 6–7 if $i < i_{max}$ else stop.
9. Scrambled LL sub-band of the watermark image for all channels (R, G, and B) was embedded into their respective LL sub-band channels (R, G, and B) of the host image using equation

$$z = (1-\text{alpha}) * x + \text{alpha} * y \tag{2}$$

where alpha is the visibility coefficients whose value can be set from 0 to 1, x is the LL coefficients of the colour channels of the host image and y is the LL coefficients of the colour channels of the watermark image and z is the resultant LL coefficients of the watermark embedded host image.

10. Two-dimensional Daubechies inverse discrete wavelet transform is applied on the resultant LL sub-band coefficients along with LH, HL, and HH sub-band coefficients of the host image.
11. Concatenate all the colour channels to form the resultant watermark embedded image.

The watermarked embedded image I_{emd} is obtained and the process of watermark embedding is completed. The resultant image contents watermark whose visibility can be controlled using parameter alpha. The colour features are invariant with respect to translation and rotation of the image. Arnold transform is commonly employed in two-dimensional domains for scattering of the pixels in an image. Thus, both provide the much-needed robustness to algorithm.

2.2 Decryption Algorithm

Let I_{emd} and I represent the 8-bit colour watermark embedded input and host image of the size $m \times n$, respectively. The steps involved in the process of watermark extraction from the host image are as follows.

1. Read the watermark-embedded input and host images of the size $m \times n$.
2. Separate the R, G, and B colour channels for both the host and watermark embedded images.
3. Each colour channel R, G, and B of the watermark embedded image was decomposed into four sub-bands LL, LH, HL, and HH using Debauches two-dimensional discrete wavelet transform.
4. Similarly each colour channel R, G, and B of the host image was decomposed into four sub-bands LL, LH, HL, and HH using Debauches two-dimensional discrete wavelet transform.
5. Scrambled LL sub-band of the watermark embedded image for all channels (R, G, and B) was used to extract scrambled LL sub-band channels (R, G, and B) of the watermark image using equation

$$y = z/\text{alpha} + x-(x/\text{alpha}) \tag{3}$$

6. Determine the number of iterations i_{max} and initialized counter $i = 0$ for Arnold transform.
7. LL sub-band of the watermark image for all channels (R, G and B) was descrambled using Arnold transform through equation

Table 1 Parameters without attack

Parameters	Lena	Baboon	Flower	Fruits
PSNR (dB)	38.16	38.16	38.16	38.16
NCC	0.573	0.572	0.573	0.575

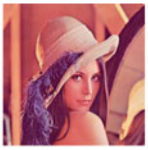
$$\left[m', n'\right] = [2, -1; -1, 1]\,[m, n] \mod 2 \tag{4}$$

8. Increment the counter $i = i + 1$.
9. Repeat steps 7–8 if $i < i_{max}$ else stop.
10. Two-dimensional Debauches inverse discrete wavelet transform is applied on the resultant LL sub-band coefficients of all colour channels.
11. Concatenate all the colour channels to form the resultant watermark image.

The complete process of embedding and extraction of watermark is discussed using colour features and Arnold transform. The watermark image obtained after extraction process was used to measure the watermarking parameters such imperceptibility, security, and robustness to common attacks such as noise, compression, and geometric attack.

3 Experimental Results

In this section, the qualitative and quantitative performance of the adaptive technique through colour features and Arnold transform is presented. Results for the four images containing varying amount of colours are provided for evaluation of the algorithm. Figure 1 shows the various host images and watermark image used for demonstration. Also, each host image was selected to determine the capability of the algorithm for invisible watermarking. Image lena has rich mix of colours which is distributed overall within the image. Image baboon has no visible colour except black and white. Whereas, images fruits and flower are rich in colours with less distribution within the images or concentrated in small areas. Peak signal-to-noise ratio (PSNR) and normalized cross correlation coefficients (NCC) are used to compute the imperceptibility and robustness of the algorithm. NCC also indicates the correlation between original and retrieved watermark images. Furthermore, algorithm embeds the watermark in colour features to promote robustness to most common attacks such as noise, geometric and compression attack. To evaluate the robustness, we employed geometrical attack such as rotation, noise attack such as salt and pepper noise and JPEG compression attack. Figures 2 and 3 shows the respective extracted watermark from lena, baboon, flower and fruits images without and with attack, respectively. Tables 1 and 2 list the measured parameters PSNR and NCC for the retrieved watermark image without and with attack, respectively.



Lena

Baboon

Flower

Fruits

(a) Host images

(b) Watermark image

Fig. 1 Images

Table 2 Parameters with attack

Attacks	Parameters	Lena	Flower	Fruits
Noise	PSNR (dB)	32.74	33.34	31.83
	NCC	0.509	0.510	0.514
Geometric	PSNR (dB)	38.16	38.16	38.16
	NCC	0.573	0.573	0.575
Compression	PSNR (dB)	37.54	33.32	32.01
	NCC	0.536	0.446	0.457

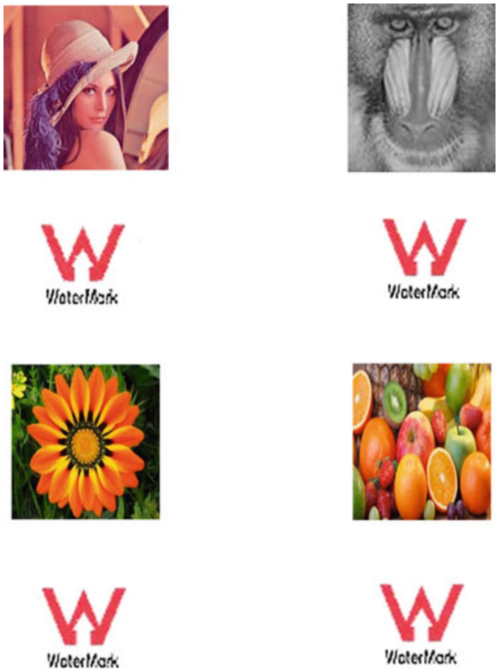


Fig. 2 Extracted watermark image from respective host images without attack

(a) Noise attack

(b) Geometric attack

(c) Compression attack

Fig. 3 Extracted watermark image from respective host images with attack

4 Conclusion

In this paper, we proposed adaptive digital image watermarking technique through colour features and Arnold transform in wavelet domain. Results clearly show the host images with different types of colour compositions achieve better masking capabilities. Also, the value of PSNR is greater than 30 dB indicates higher robustness after most common attacks. The colour features are invariant with respect to translation and rotation of the image. Images with overall colour distribution are more suitable for adaptive colour watermarking technique. Comparatively, it is understood that the value of NCC is not on higher side due to only LL sub-band coefficients were used to generate watermark image during extraction process. Further, it is required to consider all sub-band coefficients to achieve improvements in NCC. The simulation was done using MATLAB software on Intel i5 processor with 3 GHz speed and 4 GB RAM.

References

1. Andalibi, M., Chandler, D.M.: Digital image watermarking via adaptive logo texturization. IEEE Trans. Image Process. **24**(12), 5060–5073 (2015)
2. Huang, X., Zhao, S.: An adaptive digital image watermarking algorithm based on morphological Haar wavelet transform. In: International Conference on Solid State Devices and Materials Science. Phys. Procedia **25**, 568–575 (2012). (Elsevier)
3. Vaidya, P., et al.: Adaptive digital watermarking for copyright protection of digital images in wavelet domain. In: 2nd International Symposium on Computer Vision & Internet. Procedia Comput. Sci. **58**, 233–240 (2015). (Elsevier)
4. Chen, L., Zhao, J.: Adaptive digital watermarking using RDWT and SVD. In: IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE) (2015)
5. Roldan, L.R., Hernández, M.C., Chao, J., Miyatake, M.N., Meana, H.P.: Watermarking-based color image authentication with detection and recovery capability. IEEE Lat. Am. Trans. **14**(2), 1050–1057 (2016)
6. Roy, A., Maiti, A.K., Ghosh, K.: A perception based color image adaptive watermarking scheme in YCbCr space. In: 2nd IEEE International Conference on Signal Processing and Integrated Networks (SPIN) (2015)
7. Yadav, N., Singh, K.: Transform domain robust image-adaptive watermarking: prevalent techniques and their evaluation. In: IEEE International Conference on Computing, Communication and Automation (2015)
8. Shukla, D., Tiwari, N., Dubey, D.: Survey on digital watermarking techniques. Int. J. Signal Process. Image Process. Pattern Recognit. **9**(1), 239–244 (2016)
9. Tao, H., et al.: Robust image watermarking theories and techniques: a review. J. Appl. Res. Technol. **12**, 122–138 (2014)
10. Maity, H.K., Maity, S.P.: Multiple predictors based RW scheme with adaptive image partitioning. In: IEEE International Conference on Advances in Computing, Communications and Informatics (2015)
11. Siddaraju, P.M., et al.: Digital image watermarking techniques: a review. Int. J. Comput. Sci. Secur. **9**(3), 140–156 (2015)
12. Maiorana, E., et al.: High-capacity watermarking of high dynamic range images. EURASIP J. Image Video Process. (2016). (Springer)

13. Gaur, S., Srivastava, V.K.: A hybrid RDWT-DCT and SVD based digital image watermarking scheme using Arnold transform. In: IEEE 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India (2017)
14. Khalilian, H., Bajic, I.V.: Video watermarking with empirical PCA-based decoding. IEEE Trans. Image Process. **22**(4), 4825–4840 (2013)

Improved Version of Tone-Mapped Quality Index

Tushar Mane and S. S. Tamboli

Abstract High Dynamic Range (HDR) images were evolved to display smallest details of the captured image with high standards. To display HDR images on Low Dynamic Range (LDR) monitors compression is required, which is done by Tone Mapping Operators (TMOs). Recently, there are a lot of tone mapping algorithms that are available in market. Different TMO creates images with different quality. To measure the quality of such images Tone-Mapped Quality Index was proposed (TMQI). TMQI mainly depends on the two parameters. The first is structural fidelity (SF) which is very similar to structural similarity and the second, is statistical naturalness (SN). The limitation of TMQI-1 is some parameter is image independent described in below sections so, improved model TMQI-2 is proposed in this paper. In order to further improve the quality of image, iterative optimization algorithm is used. Our experimental results show that TMQI-2 is better than earlier TMQI. Further, iterative optimization increases the overall quality of image.

Keywords High dynamic range image · Structural fidelity · Statistical naturalness · Tone mapping operator

1 Introduction

In today's digital world, the importance of high-quality image has been increased. Compared to the normal images, the dynamic range of HDR images is very high. The standard LDR display is unable to cover the vast scale of HDR images. So, to display HDR images on such displays TMOs are used. Researchers worked [1, 2] to find the best TMO among the TMOs available in recent times. Few of them rely on subjective evaluation which is uneconomical, time taking, and not ideal for optimization. TMQI is a totally objective assessment model which determine image quality using parameters like statistical naturalness and structural fidelity. An

T. Mane (✉) · S. S. Tamboli
Department of E&Tc, ADCET, Ashta, India
e-mail: tushar1028@hotmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_82

improved model of statistical naturalness and structural fidelity is proposed in TMQI-2. The single quality score of TMQI-2 is better than previous TMQI as proved in the following sections. In addition, iterative optimization algorithm is described which helps to increase the quality of image.

2 Previous Work

Previously, human subjective method [3] is used to define the image quality. Subjective score depends on vision of particular subject, and his ability to spot those changes in reference image and test image, specifications of display monitor, angle, and distance between display device and human subject. Due to these factors, there was need of automatic algorithm which can define the quality of images having different dynamic ranges. Conventional method like signal-to-noise ratio and structural similarity [4] cannot be used here as they expect both images should have equal dynamic range. In [5], an attempt is made to forecast the changes between two HDR images using human visual system. But it considers HDR images having equal dynamic range.

Tone-mapped quality index of an image was proposed in [6]. It evaluates grade of image using SF and SN. The TMO used are Reinhard's method [7], Drago's method [8], Durand and Dorsey's method [9], and Mantiuk's method [10]. TMQI defines the quality of image between 0 and 1, where 1 is the highest quality. In [11], TMQI is validated using subjective score with help of rank order coefficient. It also describes parameter tuning algorithm as an application. Here, the advance algorithm of TMQI is defined which overcomes the limitations of previous version of TMQI.

3 Improved Version of Tone-Mapped Quality Index

Let G is HDR and H be LDR image. TMQI equation is described as

$$\mathrm{TMQI}(G, H) = b[S(G, H)]^{\alpha} + (1 - b)[N(H)]^{\beta} \tag{1}$$

where N is statistical naturalness, S is structural fidelity. Variables α and β along with b are described in [6].

The SF of image is calculated using sliding window algorithm over whole image. Local SF is calculated with the below equation where p is local patch in HDR and q be local image patch from LDR images.

$$S_{\mathrm{local}}(p,\ q) = \frac{2\,\sigma'_p \sigma'_q + D1}{{\sigma'_p}^2 + {\sigma'_q}^2 + D1} \cdot \frac{\sigma_{pq} + D2}{\sigma_p\, \sigma_q + D2} \tag{2}$$

where σ_p, σ_q is local standard deviations and σ_{pq} is cross correlation of two patches in LDR and HDR images. Positive constants D1 and D2 are used for stabilization. Average of all local SF measure is calculated to define overall SF

$$S_l = \frac{1}{N_l}\sum_{i=1}^{N_l} S_{local}\left(p_i,\ q_i\right) \tag{3}$$

where p_i, q_i are ith patches in the HDR and LDR images being compared, respectively, and N_ι is number of patches in *l*th scale. Local std σ in TMQI-1 is given by

$$\sigma' = \frac{1}{\sqrt{2\pi}\theta_\sigma}\int_{-\infty}^{\sigma} \exp\left[-\frac{(t-\tau_\sigma)^2}{2\theta_\sigma^2}\right]dt \tag{4}$$

where $\theta_\sigma = \tau_\sigma/3$ and threshold τ_σ are determined by contrast sensitivity function. This equation has two limitations. First, small noise such as from camera can lead to major σ value. Second, only one τ_σ value is not ideal for different images. To solve this problem Weber's law is used and σ/μ is chooses. The gradient ascent algorithm [12] is used to further enhance SF of image Q_k using Kth iterations. For this gradient of S(P, Q) is computed concerning Q and update image by

$$\widehat{Q}k = Q_k + \psi \nabla_q S(P, Q)\big|_{Q = Q_k} \tag{5}$$

where ψ is step size. For calculation of gradient ∇_Y S(X, Y), the local SF Eq. (2) can be rewritten as

$$S_{local}(p, q) = \frac{E_1\, E_2}{F_1\, F_2} \tag{6}$$

where

$$E_1 = 2\tilde{\sigma}_p\tilde{\sigma}_q + D1, \quad F_1 = \tilde{\sigma}_p^2 + \tilde{\sigma}_q^2 + D1, \quad E_2 = \sigma_{pq} + D2, F_2 = \sigma_p\sigma_q + D2 \tag{7}$$

Local SF estimate gradient relating to *q* can be expressed as

$$\nabla_Q S_{local}(p, q) = \frac{\left(E_1' E_2 + E_1 E_2'\right)}{F_1 F_2} - \frac{\left(F_1' F_2 + F_1 F_2'\right) E_1 E_2}{(F_1 F_2)^2} \tag{8}$$

where

$$E_1' = \nabla_q E_1, \quad F_1' = \nabla_q F_1, \quad E_2' = \nabla_q E_2, \quad F_2' = \nabla_q F_2 \tag{9}$$

Ultimately, gradient of overall SF estimate regarding to tone-mapped image Q is given by adding all local gradients measures

$$\nabla_Q S(P, Q) = \frac{1}{M} \sum_{i=1}^{M} R_i^T \nabla_q S_{local}(p, q)\Big|_{p=p_i,\ q=q_i} \tag{10}$$

where R_i is operator which takes *i*th local patch of image, $p_i = R_i(P)$ and $q_i = R_i(Q)$ are *i*th image patches and patch is place backed into compatible position in image by R_i^T.

In statistical naturalness, the value of P_m and P_d considered are totally image independent. In TMQI-1, the value of μ and σ are considered as 116 and 65, respectively [11]. But these values differ for different images. In TMQI-2, the overall luminance μ_e and global contrast σ_e are estimated directly from HDR image.

$$\mu_e = \frac{Z}{|x|} \left(\sum_{i,j} \frac{X^s(i,j)}{1 + X^s(i,j)} \right) \tag{11}$$

$$\sigma_e = \frac{1}{|x| - 1} \sum_{i,j} \left(\frac{X^s(i,j)}{1 + X^s(i,j)} Z - \mu_e \right)^{\frac{1}{2}} \tag{12}$$

where Z be dynamic range of tone-mapped image. Greater luminance value is decreased by scale X^s.

$$X^s(i,j) = \frac{k}{L_x} X(i,j) \tag{13}$$

where *k* is luminance level related quantity typically set between 0.09 and 0.36 for an HDR image with normal luminance level. Logarithmic equation is used to shrink greater luminance and raise small luminance values. It is given by

$$Z_x = \exp\left(\frac{1}{|X|} \sum\nolimits_{i,j} \log(\in + X(i,j)) \right) \tag{14}$$

where |X| be cardinality, $\in$ be the constant to bypass instability, and X (i, j) be luminance of HDR image at position (i, j). μ_e and σ_e are the approximate values of μ and σ. For any image, there is span of μ and σ values near μ_e and σ_e where value of SN does not differ. From this, the lower values μ_l, σ_l and upper values μ_r, σ_r are calculated. Gaussian cumulative distribution function is used to limit upper and lower bounds.

$$P_m = \begin{cases} \frac{1}{\sqrt{2\pi}\theta_1} \int_{-\infty}^{\mu} \exp\left[-\frac{(t - \tau_1)^2}{2\theta_2^2} \right] dt & \mu \le \mu_e \\ \frac{1}{\sqrt{2\pi}\theta_2} \int_{-\infty}^{2\mu_r - \mu} \exp\left[-\frac{(t - \tau_2)^2}{2\theta_2^2} \right] dt & \mu > \mu_e \end{cases} \tag{15}$$

where τ_1 and θ_1 are decided by two points (μ_l, 0.01) and (μ_e, 1). Similarly, τ_2 and θ_2 are uniquely determined by two points (μ_r, 0.01) and (μ_e, 1). The likelihood of a tone-mapped image to be natural given its std σ is computed by

$$P_d = \begin{cases} \frac{1}{\sqrt{2\pi}\theta_3} \int_{-\infty}^{\sigma} \exp\left[-\frac{(t-\tau_3)^2}{2\theta_3^2}\right] dt & \sigma \leq \sigma_e \\ \frac{1}{\sqrt{2\pi}\theta_4} \int_{-\infty}^{2\sigma_r} \exp\left[-\frac{(t-\tau_4)^2}{2\theta_4^2}\right] dt & \sigma > \sigma_e \end{cases} \tag{16}$$

where τ_3 and θ_3 are uniquely determined by two points (σ_l, 0.01) and (σ_e, 1). Similarly, τ_4 and θ_4 are uniquely determined by two points (σ_r, 0.01) and (σ_e, 1). These two quantities are multiplied to obtain overall statistical naturalness, which lies in [0, 1].

$$N(X,\ Y) = P_m P_d \tag{17}$$

To improve SN with iterative algorithm, three segment equipartition monotonic piece-wise linear function is used. A point-wise intensity transformation along with variables v and w [13], where $0 \leq v \leq w \leq Z$ selected in order that μ and σ of $Q_{k+1} = \left\{q_{k+1}^i \text{ for all } i\right\}$ approaches μ_e and σ_e of appropriate tone-mapped image is given below.

$$q_{k+1}^i = \begin{cases} (3/Z)v\,\hat{q}_k^i & 0 \leq q_k^i \leq Z/3 \\ (3/Z)(w-v)\hat{q}_k^i + (2v-w) & Z/3 \leq q_k^i \leq 2Z/3 \\ (3/Z)(Z-w)\hat{q}_k^i + (3w-2Z) & 2Z/3 \leq \hat{q}_k^i \leq Z \end{cases} \tag{18}$$

4 Experimental Results

For the experiment, ten HDR images were taken. They are converted into LDR image by using seven different TMOs. These are Ashikhmin, Drago, Durand, Fattal, Mantiuk, Pattanaik, and Reinhard TMO. For each image, SF, SN, and overall quality Q is calculated and average of it is stated in Table 1. In Table 2, initial image quality obtained from TMQI-2 and final image quality after applying iterative algorithm is listed. Constants and user-defined parameter values are taken from [11].

Table 1 shows that quality of images improved in TMQI-2 than TMQI-1 and Table 2 prove that iterative algorithm further improves the quality of image. As showed in Fig. 1, it is clear that at each iteration quality, if image increases. Initially, the Q = 0.8161, after 10 iteration Q = 0.9072, after 20 iterations Q = 0.9583, after 30 iterations Q = 0.9747, and after 40 iterations Q = 0.9787.

Table 1 Comparison between TMQI-1 and TMQI-2

Image name	TMQI-1			TMQI-2		
	SF	SN	Q	SF	SN	Q
Apartment	0.1402	0.0717	0.4499	0.6752	0.3676	0.7838
Rooftop	0.1228	0.0573	0.4297	0.6760	0.3107	0.7758
Bridge	0.2050	0.1703	0.5289	0.7271	0.2795	0.7880
Nyany	0.3794	0.3296	0.6536	0.7900	0.3648	0.8195
Lobby	0.3198	0.3007	0.6101	0.6926	0.3618	0.7861
ProvWash	0.2030	0.1071	0.5021	0.6435	0.3364	0.7622
Road	0.2608	0.4237	0.6060	0.8430	0.4051	0.8520
Snow	0.2589	0.3364	0.5954	0.7348	0.3538	0.8083
Forest	0.4225	0.3973	0.6740	0.8310	0.2597	0.8131
Door	0.2395	0.1889	0.5461	0.7181	0.3421	0.7910

Table 2 Overall quality of images before and after iterations

Image name	Initial image (Q)	Final image (Q)
Apartment	0.8253	0.9498
Rooftop	0.8274	0.9590
Bridge	0.8127	0.9666
Nyany	0.8489	0.9031
Lobby	0.8282	0.8699
ProvWash	0.8195	0.9811
Road	0.8674	0.9740
Snow	0.8382	0.9573
Forest	0.8337	0.8452
Door	0.8306	0.9568

The graph of overall quality versus No. of iteration is shown in Fig. 2a. The iterations are carried out till convergence. The “snow” image and its iteration data using 7 TMOs are considered. Figure 2b shows the ten images considered for evaluation are shown. These HDR images are tone-mapped with Reinhard’s method. These images are resized are cropped and named in table from left to right and top to bottom.

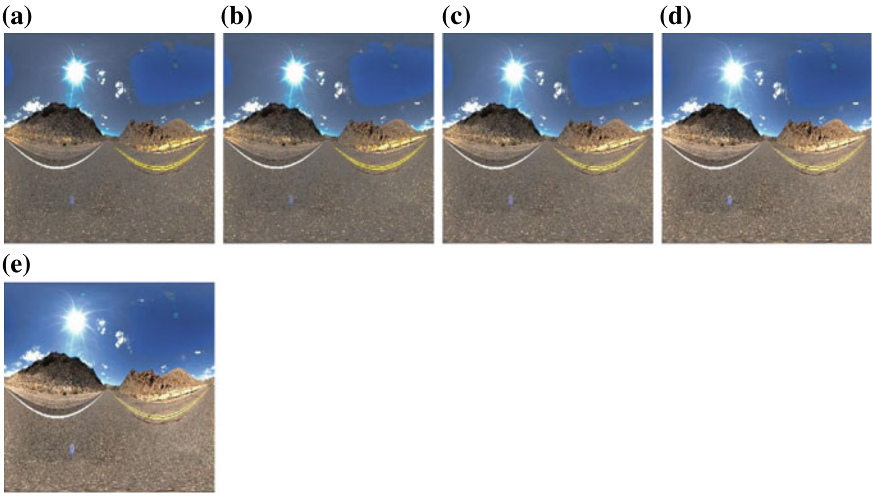


Fig. 1 HDR image tone-mapped using Durand method. Iterative algorithm is applied after TMQI-2 method **a** Initial image **b** After 10 iterations **c** After 20 iterations **d** After 30 iterations **e** After 40 iterations

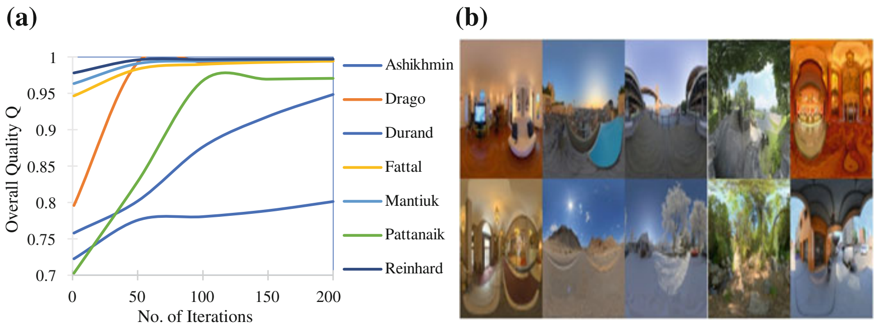


Fig. 2 **a** Graph of No. of iteration versus Q of "Snow" image. **b** Images used for experiment

5 Conclusion and Future Scope

From the experiments result, it proves that TMQI-2 performs better than TMQI-1. Both SF and SN increases and as a result of it, the overall quality of image is increased. The results show average 30% increase in quality of TMQI-2 compared with TMQI-1. The iterative algorithm further improves the image quality with average of 10%. In the future, different optimization algorithms can be developed to reduce the number of iterations. Other parameters like SNR, PSNR, structural similarity index which can be used to define the quality of image.

References

1. Ledda, P., Chalmers, A., Troscianko, T., Seetzen, H.: Evaluation of tone mapping operators using a high dynamic range display. ACM Trans. Graph. **24**(3), 640–648 (2005)
2. Čadík, M., Wimmer, M., Neumann, L., Artusi, A.: Image attributes and quality for evaluation of tone mapping operators. In: Proceedings of 14th Pacific Conference on Computer Graphics and Applications, pp. 35–44 (2006)
3. Cadík, M., Wimmer, M., Neumann, L., Artusi, A.: Evaluation of HDR tone mapping methods using essential perceptual attributes. Comput. Graph. **32**(3), 330–349 (2008)
4. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
5. Mantiuk, R., Daly, S., Myszkowski, K., Seidel, S.: Predicting visible differences in high dynamic range images—model and its calibration. Proc. SPIE **5666**, 204–214 (2005)
6. Mane, T.D., Tamboli, S.S.: Evaluating grade of tone mapped high dynamic range images. In: Proceedings of IEEE International Conference on Signal Processing and Communication (2017)
7. Reinhard, E., Stark, M., Shirley, P., Ferwerda, J.: Photographic tone reproduction for digital images. ACM Trans. Graph. **21**(3), 267–276 (2002)
8. Drago, F., Myszkowski, K., Annen, T., Chiba, N.: Adaptive logarithmic mapping for displaying high contrast scenes. Comput. Graph. Forum **22**(3), 419–426 (2003)
9. Durand, F., Dorsey, J.: Fast bilateral filtering for the display of high-dynamic-range images. ACM Trans. Graph. **21**(3), 257–266 (2002)
10. Mantiuk, R., Daly, S., Kerofsky, L.: Display adaptive tone mapping. ACM Trans. Graph. **27**(3), Art. ID 68 (2008)
11. Yeganeh, H., Wang, Z.: Objective quality assessment of tone-mapped images. IEEE Trans. Image Process. **22**(2), 657–667 (2013)
12. Wang, Z., Simoncelli, E.P.: Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics. Proc. SPIE **5292**, 99–108 (2004)
13. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer, New York, NY, USA (2006)

Robust Exemplar-Based Image and Video Inpainting for Object Removal and Region Filling

Ashvini V. Pinjarkar and D. J. Tuptewar

Abstract Inpainting is an art that restores old and damage image. Exemplar-based inpainting uses the patch-based approach. It uses patches to fill the target region of the image. Also, it uses simultaneously the texture synthesis and structural propagation. But after some iteration, the dropping effect of confidence term occurs in this method. The robust exemplar-based method avoids dropping effect by using robust priority function. The proposed video inpainting method is based on the robust exemplar-based inpainting algorithm using region segmentation. Our algorithm uses a robust priority function to avoid dropping effect and region segmentation to determine the adaptive patch size and reduced search region. The experimental results show the effectiveness of our method.

Keywords Image inpainting · Exemplar-based inpainting · Robust image inpainting · Robust video inpainting

1 Introduction

Image inpainting fills the damaged, missing region in an image with help of spatial information of the neighbor of same image. Nowadays, the image restoration is an important part of the digital image processing due to population of consumer digital cameras. Image inpainting is a method that restores old and damaged images. Also, it removes the object from image. Image inpainting is applied to red-eye correction [1], super resolution, and compression. There are several methods of image inpainting. Bertalmio [2] proposed a inpainting method that fills the target region with textural synthesis. This method fills the user selected region pixel by pixel. A. Criminisi introduced exemplar-based image inpainting method [3] that removes the large object

A. V. Pinjarkar (✉) · D. J. Tuptewar
MGM College of Engineering, Nanded, India
e-mail: ashvinipinjarkar@gmail.com

D. J. Tuptewar
e-mail: tuptewar_dj@mgmcen.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_83

from the image and fills the target region with patch approach, i.e., set of pixels. This method combines the textural synthesis and structural propagation. Robust exemplar-based inpainting using region segmentation [4] defines the robust priority function that avoids the dropping effect and uses the region segmentation that determines the adaptive patch size and reduced search region. This method gives less error propagation and more accuracy while filling the target region. So, we extend this approach to video inpainting. The proposed video inpainting method uses the robust priority function and region segmentation for selected videos.

2 Review and Background

2.1 Region Segmentation

The graph-based region segmentation algorithm [5] gives the spatial information of image. This region segmentation gives the segmentation map or segmented image that divides into number of regions. Segmentation map is given as $G=(V, E)$, where V denotes initial vertex set and E denotes a corresponding set of edges. We get segmented image through iterative merging. In each step of merging, the component of vertices C_k and C_{k+1} will merge in one segment, if the difference between two components is smaller than internal difference of two components.

2.2 Exemplar-Based Inpainting

Criminisi exemplar-based inpainting [3] removes the large object from image. It uses patch approach and iteratively searches the source region to fill the target region. This method defines the priority function in order to fill the manually selected target region of the image. This method simultaneously conducts textural synthesis and structural propagation. Priority function selects such patch which to be fill first in order to fill target region and defined as,

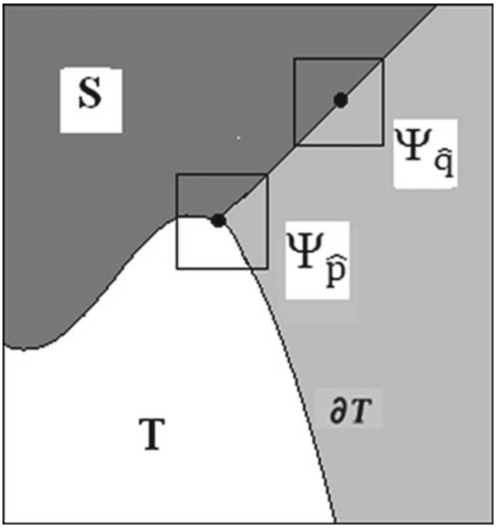
$$P(p) = C(p)D(p) \tag{1}$$

where $C(p)$ is confidence term, i.e., textural information at pixel p and $D(p)$ is data term, i.e., structural information at pixel p and given as,

$$C(p) = \frac{\sum_{q \in \Psi_p \cap S} C(q)}{\left|\Psi_p\right|},\ D(p) = \frac{\left|\nabla I_p^{\perp}.n_p\right|}{255} \tag{2}$$

where $\left|\Psi_p\right|$ is size of patch centered at p. $\nabla I_p^{\perp}$ is intensity gradient and n_p is normal vector at pixel p. 255 is normalized value for 8-bit image. This method uses fixed size

Fig. 1 Notification diagram



patch. But the error propagation is occurring in this method due to dropping effect of confidence term after number of iterations. The confidence term is nothing but the textual information from the source region and in the same way, the data term is structural information about contour, lines at pixel p. In this way, this method selects the priority patch with confidence term and data term and fills the target region. After filling complete target region, this method gives the inpainted image.

2.3 Robust Exemplar-Based Inpainting Algorithm Using Region Segmentation

Robust exemplar-based inpainting algorithm using region segmentation [4] is extension of robust algorithm for exemplar-based image inpainting [6]. This method [4] used the robust priority function and region segmentation.

Figure 1 shows target region T, source region S, boundary of target region ∂T, priority patch $\Psi_{\hat{p}}$, and source patch $\Psi_{\hat{q}}$, This method avoids the dropping effect of confidence term by using robust priority function which is defined as

$$RP(p) = \alpha . R_C(p) + \beta . D(p),\ 0 \leq \alpha, \beta \leq 1 \qquad \alpha + \beta = 1 \tag{3}$$

where regularized confidence term is given as

$$R_C(p) = (1 - \omega) . C(p) + \omega \tag{4}$$

where α and β are adaptive weighting parameters determined by DOG. ω is set to 0.7. Once priority patch is decided, we have to search the source patch in candidate source region which has minimum distance from target patch, i.e., priority patch as follows:

$$\Psi_{\hat{q}} = avg\,min_{\,\Psi_{\hat{q}} \in S} d\left(\Psi_{\hat{p}}, \Psi_q\right) \tag{5}$$

where $d\left(\Psi_{\hat{p}}, \Psi_q\right)$ is sum of square differences. Finally, the source patch is copied to the priority, i.e., target patch is as follows:

$$\Psi_{\hat{p}}(r) = \begin{cases} \Psi_{\hat{q}}(s) & \forall r \in \Psi_{\hat{p}} \cap T \\ \frac{\Psi_{\hat{p}}(r) + \Psi_{\hat{q}}(s)}{2} & \forall r \in \Psi_{\hat{q}} \cap S \end{cases} \tag{6}$$

where r and s are target patch pixels and source patch co-located pixels, respectively. This method use adaptive patch size varies from 9×9 to 17×17 depending on source region available in target patch and reduces the search region by using region segmentation.

3 Proposed Robust Video Inpainting Algorithm

Proposed video inpainting method is based on robust exemplar-based inpainting algorithm using region segmentation [4]. The proposed method uses the robust priority function defined in [6] that avoids the dropping effect due to which error propagation is very less. Also, region segmentation is used in our proposed video inpainting to reduce search region and adaptive patch size. Figure 2 shows the flow diagram of our proposed algorithm of video inpainting. In our algorithm, first, we separate number of frames from selected video. As video is converted into number of frames, each method goes through all steps one by one. Now, first frame will go to first foreground extraction. After that, frame goes through region segmentation [5]. In this step, we get segmented image which is divided into number of segmented regions.

Now, the next step is robust exemplar-based inpainting [4]. In this step, we use robust priority function defined in [6]. This method avoids the dropping effect of confidence term. Due to which, we fill the target region in more accurate manner than previous methods. Also, due to use of region segmentation in our method, we determine the adaptive patch size. The size of target patch changes, if that target patch belongs to only one segment of source region. Also, our method reduces search region, i.e., candidate source region approach is used in our algorithm. After this step, we get first inpainted frame. In this way, all these steps will repeat for each frame of the video. Finally, we get all inpainted frames that have to convert into video and we get inpainted video at output of our algorithm.

Steps to implement proposed algorithm:

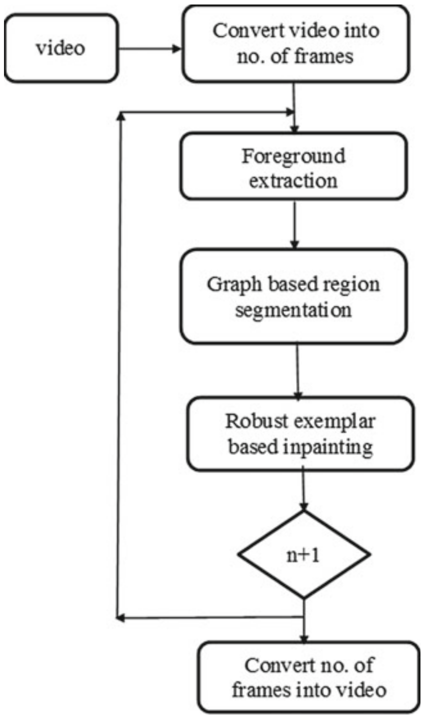


Fig. 2 Flow diagram of the proposed algorithm

1. Select the input video.
2. Convert video into number of frames.
3. Separate foreground and background of frame.
4. Apply region segmentation.
5. Apply robust exemplar based inpainting.
   a. Use robust priority function with adaptive weighting parameter.
   b. Determine adaptive patch and reduce search region.
   c. Apply Bezier curve connection method.
6. Repeat steps 3, 4, and 5 till all frames of video are processed.
7. Convert all inpainted frames into video.

Bezier curve method [7] is curve connection method that is used to fill the curves in each frame. In our method, quadratic Bezier curve connection is used.

4 Experimental Results

Here, in proposed method, MATLAB is used for experimental results. The following are the performance measures that are used for analysis.

Table 1 PSNR and RMSE result for Criminisi method and robust exemplar method

Images	Criminisi method		Robust exemplar method	
	PSNR	RMSE	PSNR	RMSE
Underwater Image1	21.7458	20.9388	35.6844	4.27047
Underwater Image2	23.3385	17.4307	37.3001	3.49331
Image1	22.7977	18.5505	36.7355	3.72788
Image2	22.6397	18.891	36.5809	3.79485
Image3	23.6799	16.7532	37.6187	3.36749
Aerial image	25.0137	14.3732	38.9279	2.8963

1. PSNR

To check the error propagation, PSNR is used which is given as

$$PSNR = log_{1o}\left(\frac{255^2}{MSE}\right) \tag{7}$$

where

$$MSE = \frac{1}{m \times n}\sum_{i=0}^{m-1}\sum_{i=0}^{n-1}[I(i, j) - K(i, j)]^2 \tag{8}$$

where $m \times n$ is size of the image and $I(i, j)$ is the original image and $K(i, j)$ is reconstructed image.

2. RMSE

Root mean square error which is given as

$$RMSE = \sqrt{MSE} \tag{9}$$

For the experimental result, we have used several types of images such as underwater images, color images, and aerial images.

Figure 3 shows the result of robust exemplar method for underwater image, color image, and aerial image, respectively. We have calculated PSNR (peak signal-to-noise ratio) which is measured in dB and RMSE for Criminisi method and robust exemplar method as shown in Table 1. This shows that the performance of the robust exemplar method is greater than Criminisi method. These performance measures are used here to check the quality of inpainted image for both methods. PSNR is higher for the robust exemplar method than the Criminisi method. But the time required for this method is greater than Criminisi method as shown in Table 2.

(a) Input images

(b) Output images

Fig. 3 Result of robust exemplar method for underwater image, image1 and aerial image

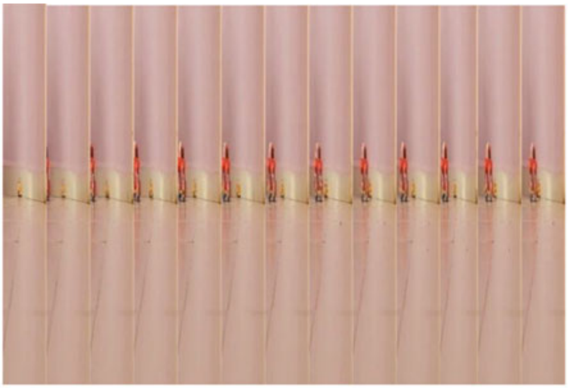
Table 2 Total time required for Criminisi method and robust exemplar method

Images	Total time required			
	Criminisi method		Robust exemplar method	
Underwater Image1	21.7458	20.9388	35.6844	4.27047
Underwater Image2	23.3385	17.4307	37.3001	3.49331
Image1	22.7977	18.5505	36.7355	3.72788
Image2	22.6397	18.891	36.5809	3.79485
Image3	23.6799	16.7532	37.6187	3.36749
Aerial image	25.0137	14.3732	38.9279	2.8963

Then, also robust exemplar method is efficient due to more accuracy in patch filling. This gives high PSNR. So, we extend this method for video inpainting.

Figure 4 shows the results of proposed video inpainting algorithm. There are 91 frames of original video. Among 91 frames, 13 frames are shown in Fig. 4 (a) shows original video frames and (b) shows that the object, i.e., car is removed from each frame by the proposed video inpainting method. Here, each frame goes through the all steps of the proposed video inpainting algorithm. From each frame, the car is removed and region is filled with the background of the same image.

Fig. 4 Result of the proposed video inpainting algo



(a) Original video frames

(b) Inpainted video frames

As per analysis, the robust inpainting algorithm gives high PSNR than that of previous work although required computation time is more. Due to use of graph-based region segmentation, number of iteration reduces and error propagation is less for image. Hence, the proposed method is the robust video inpainting. However, the proposed video inpainting method requires more computation time. In future, an algorithm can be developed that reduces required computation time and gives high quality of video inpainting.

Referances

1. Yoo, S., Park, R.-H.: Red-eye detection and correction using inpainting in digital photographs. IEEE Trans. Consumer Electron. **55**(3), 1006–1014 (2009)
2. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the ACM Conference Computer Graphics (SIGGRAPH), pp. 417–424, New Orleans, LU, July 2000
3. Criminisi, A., Perez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. **13**(9), 1200–1212 (2004)
4. Lee, J., et al.: Robust exemplar-based inpainting algorithm using region segmentation. IEEE Trans. Consum. Electron. **58**(2) (2012)
5. Felzemszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Comput. Vis. **59**(2), 167–181 (2004)
6. Cheng, W.-H., Hsieh, C.-W., Lin, S.-K., Wang, C.-W., Wu, J.-L.: Robust algorithm for exemplar-based image inpainting. In: Proceedings of the International Conference on Computer Graphics, Imaging and Vision 2005, pp. 64–69, Beijing, China, July 2005
7. Hung, J.C., Huang, C.-H., Liao, Y.-C., Tang, N.C., Chen, T.-J.: Exemplar-based image inpainting base on structure construction. Int. J. Softw. **3**(8), 57–64 (2008)

Comparative Analysis for Steganographic LSB Variants

Namrata Singh and Jayati Bhardwaj

Abstract Combining the best features of steganography and cryptography is the trending concept which is being followed for the purpose of information security. Hence, this combination is making the data more powerful and secure against the prevailing security attacks and breaches. This paper represents the implementation of this combination on the two LSB variants, namely sequential LSB and randomized LSB. A comparison among the two approaches is carried out by adding a secret text into a video cover file. The concept of chaotic sequence has been used as the security approach that converts the secret data into random bits pattern. The proposed work uses the traditional LSB approach as basic steganographic model. The inference on the basis of parameters concludes that the randomized LSB shows better results than the sequential LSB scheme.

Keywords CVSS · Hash LSB · LSB + 3 · Least significant bit (LSB)

1 Introduction

With every increase in the quality of work in the field of information security, the need to find out the best possible solutions out of the proposed ones is quite imposing. Finding out the best solutions is a result of comparative analysis being carried out on two or more than two techniques related to a relevant field. Information security concerns with two broad topics: Steganography and Cryptography. Steganography deals with secrecy while cryptography deals with security. The combination of the two wide fields leads to a new concept of Metamorphic Cryptography. This metamorphic combination first secures the data through encryption and then embeds the encrypted file into the cover file. The work hereby presented in this paper uses

N. Singh · J. Bhardwaj (✉)
AKTU Lucknow, Lucknow, India
e-mail: jayatibhardwaj2@gmail.com

N. Singh
e-mail: nam2817210@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_84

Chaotic Sequence concept for the encryption purpose. These are complex, nonperiodic and random in time domain and widely used in encryption algorithms since 1989 [1]. Chaotic arrangement has also been used in the field of cryptography for providing randomness to our information. The random values created by the chaotic sequence lies in the boundary values of [0, 1]. Values collection and calculation will be performed using the following formula [2].

$$X_n + 1 = \mu * X_n * (1 - X_n) \tag{1}$$

where X_n is a number between zero and one that represents the ratio of existing population to the maximum possible population. The values of interest "μ" parameter lie in the interval [0, 4] [3].

The paper represents a comparison among two LSB techniques using chaotic sequence as the encryption technique. The LSB types being compared here are Sequential LSB and Randomized LSB. The secret data being hidden in the cover video file is a text of 350 bits. Main parameters on the basis of which the comparisons are being done are PSNR, MSE, and SNR. Also, histograms and embedded frames are also used for the defining the best LSB approach out of the two. The formulas for the deciding parameters are as follows:

$$SNR = 10 \log_{10}\left(P_{signal}/P_{noise}\right). \tag{2}$$

$$PSNR = 10 \log_{10}(255/MSE)^2. \tag{3}$$

$$MSE = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} (x_{j.k} - x'_{j,k})^2 \tag{4}$$

where P_{signal} and P_{noise} are average power of both signal and noise in Eq. (2). Equation (4) represents a noise-free $M \times N$ monochrome image x and its noisy approximation x' [4, 5].

This paper is organized in the following manner: Sect. 1 contains the introductory part describing the basic concept related to the proposed work. Section 2 contains the literature survey of all the related work concerning same domain. Section 3 contains the proposed work involving all algorithms and flowcharts. Section 4 contains the results and observations while Sect. 5 concludes the paper.

2 Literature Review

As defined in [6], text is embedded into the least significant bit of each video frame. Decoding of the secret message is done by extracting the LSB from the encoded image/frame. Stego key is used in the form of polynomial equation which increases the bits capacity of cover video frames. A modified LSB is proposed by Ramalingam [7] in the form of stego machine. This stego machine embedded the secret text message into a video file. The stego machine is platform independent which results in high portability and consistency. In a scheme proposed by Manpreet and Amandeep kaur, the combination of both steganography and cryptography features is used. The scheme derives insertion of text in a video cover with the help of hash LSB algorithm [8, 9] with a new encryption algorithm. Hash function is used to evaluate the positions where to hide the data in the cover file. RSA encryption algorithm implemented here before the embedding. The embedded video so produced is secure and conforms to the original video in quality terms. The methodology of hiding acoustic data into a video file using LSB technique is presented in the paper [10]. The proposed technique shows remarkable results as minimal changes are observed after insertion. The PSNR and MSE values are found to be high and low, respectively.

The work presented in paper [11] by Pritish et al. introduced a new video steganographic technique for high-resolution .avi files. The method changes the LSB and LSB + 3 bits in the alternate bytes of the cover file (video). The scheme possessed various advantages like high imperceptibility, security, less computational time, and increased capacity. Similarly index based on more secure video steganography is mentioned in [12]. This simplifies the decoding and extraction procedure by analyzing only selected frames as defined by the index. This reduces computational time as the extraction is done selectively instead of sequentially frame by frame. The proposed scheme claimed to be more secure with the extraction time. The comparison among LSB and Random byte hiding technique is done by Ashish and Rachna [13]. The work shows the advantage of using video as cover by overcoming the disadvantages of using image as cover. The lossy-type technique is implemented as lossless-type technique is quite complex and require more run time. The results inferred the random byte hiding technique as the better one as compared to LSB. The encryption and decryption time taken is quite less in random as compared to the LSB.

Another approach of video steganography of hiding a video into another video is proposed in paper [14]. The results are highly secure resulting in no distortion in the quality of video. Mohamed Elsadig et al. [15] used LSB technique along with 3-3-2 embedding approach for video steganography. 3-3-2 approach defined the number of bits to be stored at the RGB plains, respectively. However, the insertion is being done sequentially using LSB technique. A new video steganographic approach using LSB is proposed in [16] where security is inculcated by generating an indexed based chaotic sequence and arranging the frame pixels in accordance to it. The results show high quality of imperceptibility amongst the videos. A steganographic technique of hiding text into a video cover is mentioned in the paper by Sunil and Rajshree [17]. Comparison among different LSB approaches is done—1LSB, 2LSB, and 4LSB. The paper concludes 4LSB as an effective algorithm for the large amount of secret data hiding purpose. CVSS (Compressed Video Secure Steganography) is the new algorithmic approach defined by Bin Liu et al. [18]. The variance of each frame decides the embedding payload. Compressed domain is used for both embedding and detection. Steganalysis is done on the basis of collision-based video Steganalysis algorithms. Visually, no distortion is observed in the cover file after embedding procedure.

3 Proposed Work

3.1 Working Model

The flowcharts for both the LSB techniques are being represented in the figures. Figure 1 shows the working model of the sequential algorithm using LSB. Figure 2 shows the working model of the randomized LSB technique using video as cover and text as secret object. The two basic models of cryptography and steganography are defined clearly in the schemes. Both the models use chaotic sequence in the text for encryption before embedding through Eq. (1) as mentioned in Sect. 1. XORing among the chaotic bits and LSBs is done before insertion/embedding into the LSB place of the frames. In sequential LSB, each frame participates in the LSB replacement while in randomized technique, a crypt random generator function selects a random frame for insertion purpose. However, the cryptographic and steganographic approaches remain the same. Section 3.2 discusses the algorithm for the embedding process of both approaches.

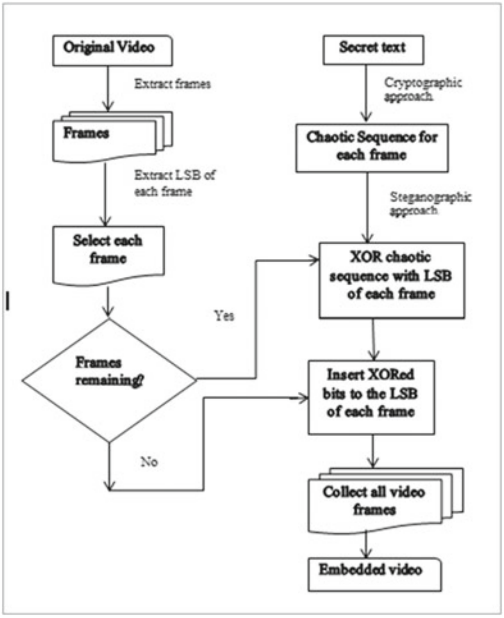


Fig. 1 Flowchart of the Sequential LSB technique

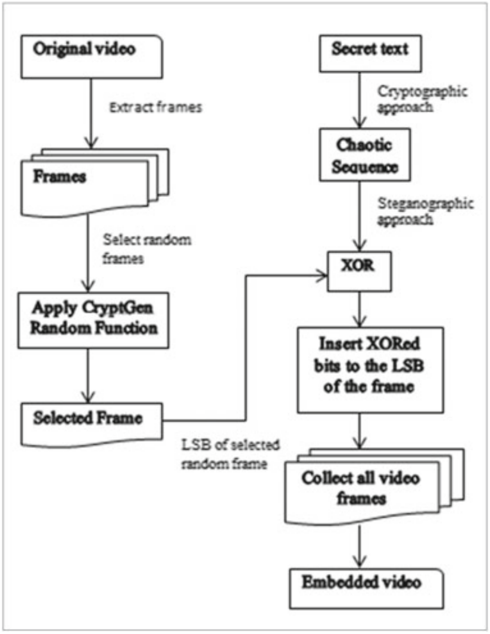


Fig. 2 Flowchart of the Randomized LSB Technique

3.2 Algorithms for Sequential LSB Coding

Embedding:

1. Read the video and text.
2. Convert text file to binary data stream.
3. Extract all video frames into variable "v".
   (a) Take the variable K for frame counter.
   (b) Read the video file.
   (c) Keep looping for all frames.
   (d) Compile all frames in the 4th dimension of variable "v".
   (e) The total number of LSBs available in video file is found.
4. Check for space requirement for embedding.
   (a) If (reqspace > availspace), we get the error message.
   (b) If (reqspace < availspace), we simply embed the text into the video file.
5. LSB embedding
   (a) Use counter for the text bits.
   (b) Calculate the each row pixel number.
   (c) Calculate the each column Pixel number.

   (d) Select the channel (1, 2, 3) red, green or blue.
   (e) Calculate the each frame of the video.
   (f) The original pixel value converted to binary.
   (g) Apply chaotic sequence formula to the message bits
   (h) XOR chaotic sequence with the LSB of each frame
   (i) Change the LSB value to the XORed bits.
   (j) Convert to decimal.
   (k) Assign to new pixel value.
   (l) Put the coded video into this variable.

6. Write coded video in a new file to get the embedded video.

3.3 Algorithms for Random LSB Coding

Embedding Algorithm:

1. Read the video and text.
2. Convert text file to binary data stream.

   (a) Convert the text to a binary stream.
   (b) Find out the required number of LSBs to hide data.

3. Extract all video frames into a variable say "v".

   (a) Take the variable K for frame counter.
   (b) Read the video file.
   (c) Keep looping for all frames.
   (d) Compile all frames in the 4th dimension of variable "v".
   (e) The total number of LSBs available in video file is found.

4. Check for the space condition for embedding

   (a) If (reqspace > availspace), we get the error message.
   (b) If (reqspace < availspace), we simply embed the text into the video file.

5. CryptGenRandom to generate random number for frame selection.
6. LSB embedding

   (a) Use Random variable "r" containing the frame number.
   (b) Find the length of the message: m = length(message) * 8;
   (c) Convert message to binary stream.
   (d) Apply chaotic sequence formula on message bits.
   Message bits are converted into random sequences.
   XOR the random bits with the LSBs of frame.
   (e) Change the LSB value to the message bits resulted by XORing.
   (f) Write the coded video/image into the output folder.
   (g) Write coded video in a new file to get the embedded video.

4 Results and Observations

The proposed approaches are implemented on MATLAB 2016. The secret message encoded is in the form of text of 350 bits. Four different videos of different sizes are used for the comparison. Tables 1 and 2 represent the Sequential LSB and randomized LSB parameters in which the PSNR values of the random LSB vary greatly as compared to sequential one. Same goes with MSE of random LSB technique; its value is lesser in the first one than the latter. These variations symbolize the effectiveness of random LSB coding as in a good steganographic approach the PSNR values go higher while MSE values go lesser. Table 3 represents the comparison among the embedded frames and the original frames along with their histograms in both the approaches scenario. Both the original and embedded frames possess negligible differences among them and the same goes with histograms. On visual detection, no difference and distortion is observed among both frames and histograms (Table 3) .

5 Conclusion

The comparison among above two steganographic techniques shows that the randomized LSB is better than the sequential LSB technique. The PSNR values in the randomized LSB approach are comparatively higher than the traditional approach. Similarly, the MSE values are going less in more values in random LSB than the sequential (as shown in Tables 1 and 2). The Steganalysis of both approaches show Random technique as more imperceptible strong and secure. The time elapsed in embedding and extraction is also small in the randomized LSB. The video retains more originality in the same approach.

Table 1 Sequential LSB parameters

Parameters	Type	Nature.avi	Shuttle.avi	Tower.avi	Walk.avi
SNR	Original	19.3901	18.9563	10.4789	15.5013
	Embedded	20.3957	21.9636	11.5257	16.5836
PSNR	Original	20.9306	22.3623	20.6080	22.2204
	Embedded	21.9430	23.3696	20.6413	22.6813
MSE	Original	524.5850	383.7629	561.6528	389.6067
	Embedded	521.3834	381.5445	557.6314	387.5631

Table 2 Randomized LSB parameters

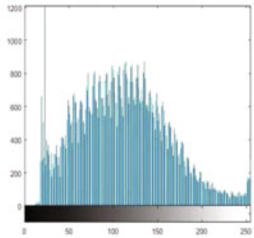
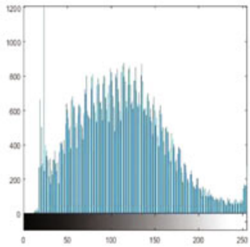
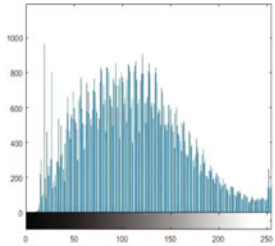
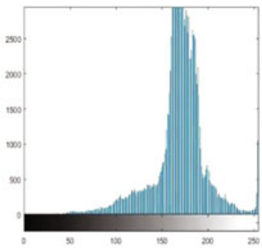
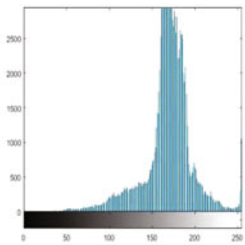
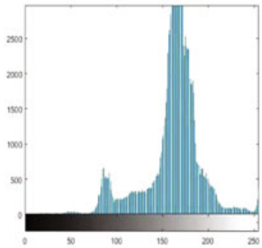
Parameters	Type	Nature.avi	Shuttle.avi	Tower.avi	Walk.avi
SNR	Original	20.9707	22.3412	20.5860	22.0057
	Embedded	22.9871	25.3504	26.6816	28.1685

(continued)

Table 2 (continued)

Parameters	Type	Nature.avi	Shuttle.avi	Tower.avi	Walk.avi
PSNR	Original	19.4321	18.5866	10.4570	15.4386
	Embedded	21.5389	23.9744	15.5657	21.6013
MSE	Original	524.0303	369.3592	556.4750	402.2987
	Embedded	520.5691	363.7040	553.9524	398.4731

Table 3 Comparison among sequential and random LSB technique on the basis of frames

Original Frames	**Sequential Approach**	**Random Approach**
		
		
		
		

References

1. https://en.wikipedia.org/wiki/Logistic_map. Accessed 20 Dec 2017
2. Trivedi, M.C., et al.: Metamorphic cryptography using strength of chaotic sequence and XORing method. J. Intell. Fuzzy Syst. **32**, 3365–3375 (2017)
3. https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio. Accessed 20 Dec 2017
4. https://en.wikipedia.org/wiki/Signal-to-noise_ratio. Accessed 20 Dec 2017
5. Ogras, H., Turk, M.: Digital image encryption scheme using chaotic sequences with a nonlinear function. World Acad. Sci. Eng. Technol. Int. J. Comput. Electr. Autom. Control Inf. Eng. **6**(7) (2012)
6. Swathi, A., Jilani, S.A.K.: Video steganography by LSB Substitution using different polynomial equations. Int. J. Comput. Eng. Res. (IJCER) **2**(5) (2012)
7. Ramalingam, M.: StegoMachine-video steganography using modified LSB algorithm. World Acad. Sci. Eng. Technol. **50** (2011)
8. Kaur, M., Kaur, Er.A.: Improved security mechanism of text in video by using steganographic technique: a review. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **4**(5) (2015)
9. Singh, N., Yadav, V.K.: Trends in digital steganography: a survey. Int. J. Comput. Appl. (2017)
10. Gosalia, S., et al.: Embedding audio inside a digital video using LSB Steganography. In: 2016 International Conference on Computing for Sustainable Global Development (INDIAcom). IEEE (2016)
11. Bhautnage, P., Jeyakumar, A., Dahatonde, A.: Advanced video steganography algorithm. Int. J. Eng. Res. Appl. (IJERA) (2013)
12. Balaji, R., Naveen, G.: Secure data transmission using video steganography. In: 2011 IEEE International Conference on Electro/Information Technology
13. Bole, A.T., Patel, R.: Steganography over video file using random byte hiding and LSB technique. In: International Conference on Computational Intelligence and Computing Research. IEEE (2012)
14. Yadav, P., Mishra, N., Sharma, S.: A secure video steganography with encryption based on LSB technique. In: International Conference on Computational Intelligence and Computing Research. IEEE (2013)
15. Elsadig, M., et al.: High rate video streaming steganography. In: Proceedings of 2009 IEEE International Conference on Future Computer and Communications (2009)
16. Paul, R., et al.: Hiding large amount of data using a new approach of video steganography. In: Confluence 2013: The Next Generation Information Technology Summit (4th International Conference) (2013)
17. Moon, S.K., Raut, R.D.: Analysis of secured video steganography using computer forensics technique for enhanced data security. In: Proceedings of 2013 IEEE 2nd International Conference on Image Information Processing (ICIIP-2013)
18. Liu, B., Liu, F., Yang, C., Sun, Y.: Secure steganography in compressed video bitstreams. In: 3rd International Conference on Availability, Reliability and Security. IEEE (2008)

Integrating Machine Learning Tool to Improve DSS Design

R. G. Joshi and H. S. Fadewar

Abstract This paper describes how a machine learning tool can be applied to decision support system. We have used fuzzy logic to enhance performance of DSS. Further, system developed is implemented in agriculture domain for selection of suitable crop. Selection of crop is complex process as it involves number of parameters where uncertainty is more common for example rainfall, suitable seeds, fertilizers, number of soil parameters, temperature, air quality, humidity, and so on. The present work focuses on soil parameters and few other parameters which support proper growth of crops. Fuzzy logic is applied to those parameters for handling data uncertainty. This is an attempt to suggest proper decision and reduce the burden by designing new DSS. Experimental set-up shows increased crop production up to 10–12%.

Keywords Machine learning · Fuzzy logic · Decision support system

1 Introduction

Machine learning is a ability to make computer automatically learn and improve the quality of solution from previous experiences [1]. There are varieties of situations handled effectively by the machine learning algorithms like speech recognition, pattern matching, and so on. There are several machine learning tools like Neural Network called Artificial Neural Network, Fuzzy Logic, and Genetic Algorithm for handling different types of programming situations. Principal component analysis and independent component analysis may are also used as an analysis tool due to their accuracy [2]. In proposed system, we have used fuzzy logic. Because we want to process certain linguistic variables and fuzzy logic is best choice for it.

R. G. Joshi · H. S. Fadewar (✉)
School of Computational Science, S.R.T.M. University, Nanded, MS, India
e-mail: fadewar_hsf@yahoo.com

R. G. Joshi
e-mail: yma.joshi@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_85

2 Decision Support System

Decision-making process is obvious part of any organizational process. Decision support system are computer-assisted system which helps the decision maker to arrive at certain decision. According to *Herbart Simon* [3], process of decision-making is divided into three parts: identification of problem domain that requires decision-making called intelligence phase, finding out solutions and their feasibility called design phase, and selection of best possible alternative called choice phase. In conventional programming, the program written once becomes rigid. To maintain flexibility in the program, here we have decided to use fuzzy logic as a machine learning tool for decision-making process. In recent development, number of technologies can be observed in DSS design process. Two important techniques under artificial intelligence are Rule-Based System (RBS) and Case-Based Reasoning (CBR) [4, 5]. RBS do not express the things in declarative and static way rather it applies set of "if-then" rules to problem under investigation at specific instances. On the contrary, in CBR techniques, expertise is recorded in library of different cases happened. Normally, each case is associated with specific problem and solution for it. The proposed system is built around RBS.

3 Steps Involved in System Development

Overall steps in system development are shown in the following diagram (Fig. 1).

3.1 Design of FIS (Fuzzy Inference System)

The first step in system development is to frame fuzzy system. Fuzzy system framework involves process of conversion of crisp value into fuzzy value [6]. Further, we have to apply fuzzy operator to antecedent, proposition from the antecedent to the consequent, aggregation of the consequents across the rules, and defuzzification. These are the important points one has to consider while developing fuzzy inference system. Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic [7]. There are two types of fuzzy inference systems, namely mamdani and sugeno type. Mamdani is more commonly used fuzzy inference system in which singleton output of membership function is expected, whereas Sugeno type of system are used when output is either linear or constant. For the proposed system, mamdani type of FIS is used.

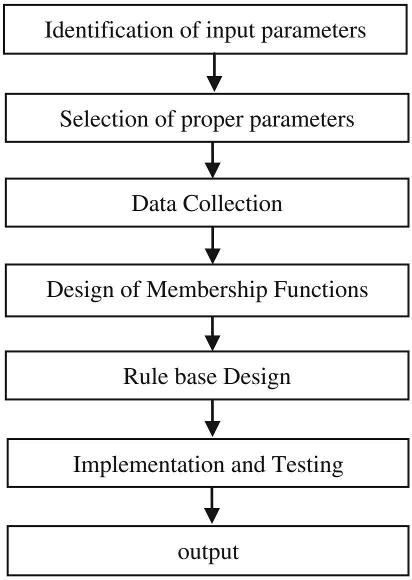


Fig. 1 Research flow diagram of fuzzy inference system

3.2 Selection of Input Parameter

As shown in Fig. 2, input given to the system is crisp or fuzzy data. If data given to the system is crisp, it must be converted into fuzzy. This process is called fuzzyfication. For development of present system, fuzzy logic toolbox from MatLab is used for the entire process. As we want to implement the proposed system in agriculture domain, by consulting agriculture expert, we have decided parameters which affect on production. This process is very complex because a user has to consider different parameters where uncertainty is more common (Fig. 3). For example, atmospheric condition, rainfall, proper selection of seeds, suitable fertilizers, soil parameters, and so on. In the proposed system, we concentrate on soil parameters. It is mandatory for user to test soil form laboratory, so that system will give accurate result. For simplicity, it was decided to consider majority of soil parameters and few closely related with it. In this way, 15 different parameters are decided and system is build around it. The selected parameters are depicted in Fig. 4.

3.3 Membership Function Design

Figure 3 shows membership function editor from MATLAB. Appropriate design of membership function for input and output variables is done through this section. This is the further important step in system development.

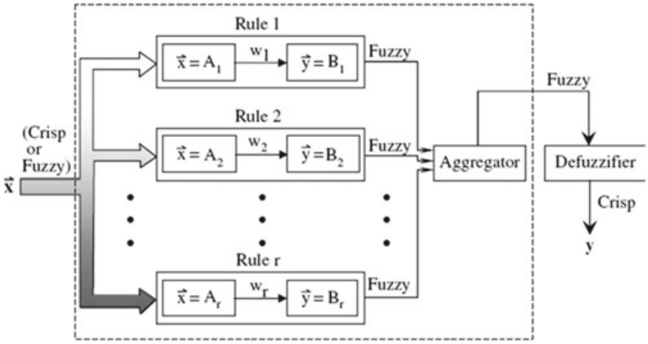


Fig. 2 Fuzzy inference system (*source* www.Researchgate.net)

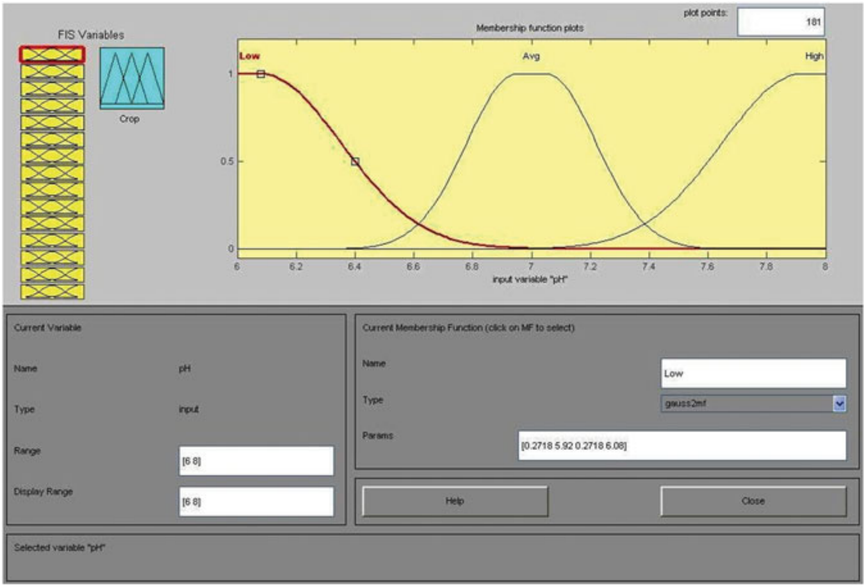


Fig. 3 Membership function editor

As per system requirement, gauss2mf (Gaussian) membership function for fuzzyfication of input and *trimf* (Triangular) membership function for fuzzyfication of output is used.

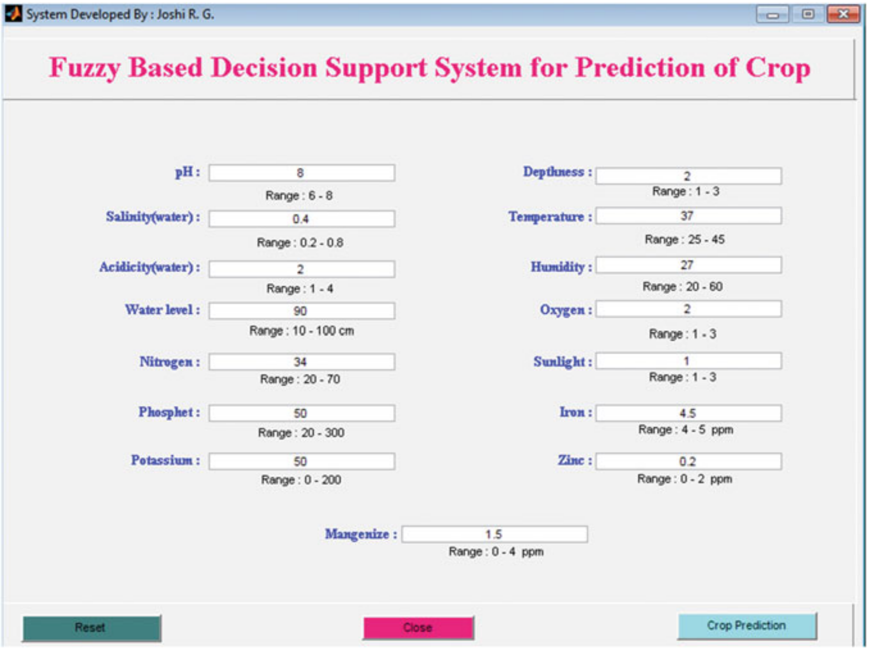


Fig. 4 Input given to system

3.4 Designing Rule-Based System

Once parameters and their membership functions are decided, the next major step in system design is to write rule-based system. Rules are defined in following manner.

If (pH is High) and (Salwater is Low) and (Acdwater is High) and (WaterLvl is High) and (Nitrogen is Vhigh) and (Phosphate is Low) and (Potessium is Low) and (Depthness is Avg) and (Temp is Avg) and (Humidity is Avg) and (Oxygen is High) and (Sunlight is Avg) and (Iron is Avg) and (Zinc is Avg) and (Maganese is Low), then (crop is Wheat). This is another example of crop Green gram.

If (pH is Avg) and (Salwater is Avg) and (Acdwater is High) and (WaterLvl is High) and (Nitrogen is Vhigh) and (Phosphate is Low) and (Potessium is Low) and (Depthness is High) and (Temp is Avg) and (Humidity is Low) and (Oxygen is High) and (Sunlight is Avg) and (Iron is Low) and (Zinc is High) and (Maganese is High), then (crop is Green gram).

In this way, 22 different rules have been defined for the proper output. Writing such rule with several combinations is lengthy process (Fig. 5).

Figure 4 illustrates the input parameters given to system and their reference values, e.g., Nitrogen, Phosphate, Potassium, Iron, Zinc, and so on. System gives output in fuzzy form and defuzzification is done by using centroid method. Finally, Fig. 6 show output generated by the system.

Fig. 5 Output of the system

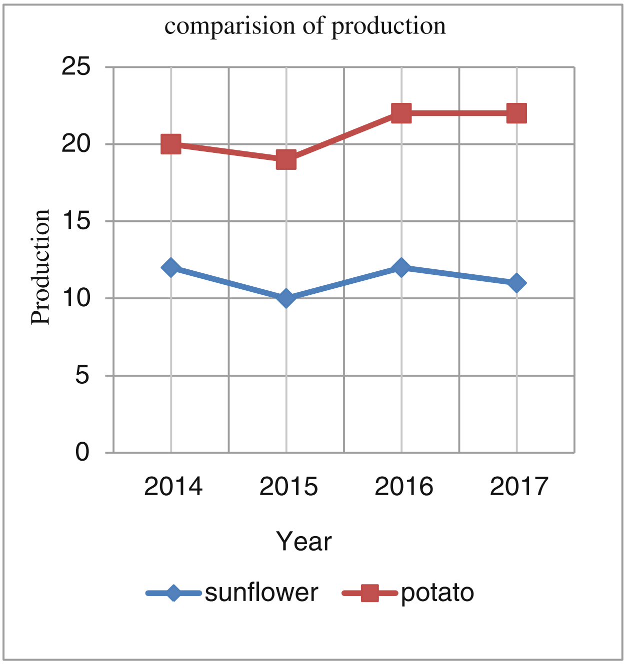


Fig. 6 Comparison of crop productivity

Table 1 Comparative study of crop production

Sr. no.	Year	Traditional crop	Production quintal/acre	System recommended crop	Production Quintal/acre	
					Old values	New values
1	2014	Sun flower	12	Potato	18	20
2	2015	Sun flower	10	Potato	17	19
3	2016	Sun flower	12	Potato	20	22
4	2017	Sun flower	11	Potato	21	22

4 Result and Discussion

Earlier researchers Sri Hartati and Imas have used fuzzy approach for evaluating land suitability and selecting crops. But they have considered 10 soil parameters, whereas we have considered 15 soil parameters [8]. Kefaya et al. [9] used adaptive fuzzy neuro approach for crop yield production. But considering only single crop, we covered 25 different crops. Mawle and Chavan [10] used fuzzy system but their focus is to predict fertility of soil, whereas we provide exact suitable crop. Kaur [11] used machine learning approach as it covers irrigation and disease management in Indian agriculture. Our system recommends suitable crop using fuzzy concept.

Traditional system for selection of crop is based on past experiences. Our experimental set-up is based on scientific soil testing. It has been observed that, system developed so far recommends other crops than the traditional system, and it shows increased growth in production.

Table 1 shows difference in production of crop from traditional system to proposed system. It clearly shows increased production of crop. Graph in Fig. 6 indicate the same. System developed around soil parameter gives suitable crop. According to agriculture expert, the parameter selected affects 10–12% on total productivity. We have tested results from traditional system and fuzzy-based system and results from developed system are found to be satisfactory and shows 10–12% increased growth.

References

1. Ahmad, I.: Fuzzy Logic For Embedded Systems Applications, pp. 113–118, Elsevier India Pvt. Ltd. (2006)
2. Patil, M., Iyer, B., Arya, R.: Performance evaluation of PCA and ICA algorithm for facial expression recognition application. In: Proceedings of Fifth International Conference on Soft Computing for Problem Solving, pp. 965–976 (2016)
3. Jawdekar, W.: Management Information System, pp. 23–47. TMH (2002)
4. Hayes-Roth, F.: Rule-based systems. Commun. ACM **28**(9), 921–932 (1985)
5. Ignizio, J.: Introduction to Expert Systems: The Development and Implementation of Rule-Based Expert Systems. McGraw-Hill, New York (1990)

6. Chakraborty, D., Pal, N.R.: A neuro-fuzzy scheme for simultaneous feature selection and fuzzy rule-based classification. IEEE Trans. Neural Netw. **15** (2004)
7. Shaofeng, et al.: Integration of Decision Support Systems to Improve Decision Support Performance. Springer, London (2009)
8. Hartati, S., Sitanggang, I.S.: Fuzzy based DSS for evaluating land suitability selecting crop. J. Comput. Sci. **6**(4), 417–424 (2010). ISSN 1549-3636
9. Qaddoum, K., Hines, E., Illiescu, D.: Adoptive neuro-fuzzy modeling for crop yield prediction. Recent Research in Artificial Intelligence, Knowledge Engineering and Database, ISBN 978–960-474-273-8 (2014)
10. Mawle, Chavan, V.: Fuzzy inference system for productivity and fertility of soil. Int. J. Eng. Dev. Res. **2**(3), 2321–9939 (2014)
11. Kaur, K.: Machine learning: application in indian agriculture. Int. J. Adv. Res. Comput. Commun. Eng. **5**(4), 2278–1021 (2016)

PSO-Based Text Summarization Approach Using Sentiment Analysis

Shrabanti Mandal, Girish Kumar Singh and Anita Pal

Abstract In the present era of technology, most of the human activities are controlled and monitored by the electronic devices and still, people are working for more advanced technology and hence to fulfill the customers requirement. Government is also promoting digitization of data which results in large volume of data. To manage digital data, some approach is required to retrieve the data efficiently. Till now, so many techniques have been proposed for retrieving data in original form as well as compact form. This paper focuses on the technique for retrieving the data (text) in compact form or summarizes form. To achieve this goal, the concept of Particle Swarm Optimization (PSO) with sentiment analysis has been used. PSO has been used in the field of text summarization and the result is remarkable. Besides PSO, Sentiment Analysis (SA) has been proved its importance in the same research field.

Keywords Information retrievals · Text summarization · Particle swarm optimization · Sentiment analysis

1 Introduction

Text summarization is a process of representing the original documents in a précis form which contains only important aspects or points not the details [1]. Some summarizer generates summaries by the computing devices, keeping its main features

S. Mandal · G. K. Singh (✉)
Department of Computer Science & Applications, Dr. Harisingh Gour Central University, Sagar, MP, India
e-mail: gkrsingh@gmail.com

S. Mandal
e-mail: shrabmandal@gmail.com

A. Pal
Department of Mathematics, National Institute of Technology, Durgapur, West Bengal, India
e-mail: anita.buie@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_86

and points is called automated summarizer [2]. The main objective of summarizer is to produce the same result as human being summarize the text. To compete with the human summaries, there are lots of challenges faced by researchers. One of the most important challenges is to choose the most valuable features and select the sentences keeping in mind not to repeat the concept and cover overall matter and not overruled the diversity constraints [3]. Depending on the input source summarization is of two types, single document summarization and multi-document summarization. As the name is single document summarization, one document is used as input. There are many methods have been implemented for single document summarization [4]. The input of the multi-document summarization is multiple source documents. Single document summarization technique is simpler than multi-document summarization technique [5]. Based on the output of summarizer, it is classified into two types, the first one is extractive and the second is abstractive. In extractive summarization, some important aspects or a part of the source have extracted and arranged to produce the summary. Rather than selecting the important parts of the document, the new sentence by using the important aspects of original document is constructed in abstractive summarization. This technique is more sophisticated and complicated then extractive technique. In the 1950s, the concept of text summarization technique has introduced and in the 1970s, artificial intelligence (AI) emerged [6]. The aim of AI is to extract the knowledge for identifying entities and relationship among entities, but there was some limitation and that leads to incomplete analysis of entities.

The basic concept of PSO, used for solving the problem of swarm intelligence, is based on the population evolutionary algorithm. Adopting the social behavior of bee flying, birds flocking, or fish schooling simulates a social simplified model for presenting the information sharing in a simple way. The best effort of the PSO is to discover the most appropriate or favorable region of the search space called population. The particle is identified as a normal member of population and group of such particle is called swarm. With a random initialized velocity, the particle moves around in the search space. The velocity of a particle is dynamic and proportional to flying experience of its own and swarms. So it can be assumed that all the particles get their best position in the search space [7]. In [8], an optimization-based unsupervised method has been proposed which has improved the efficiency of automatic text summarization. An efficient quantum-behaved PSO algorithm has been proposed in [9] which used the concept of weighted mean in best position. For text summarization, some approaches which are based on fuzzy logic helped to improve the performance. Such a model has been presented in [10] which worked on extraction strategy and selected the sentences from the source documents based on degree of importance and high relevance score. Many researchers have also worked on multi-document summarization. In [11], Y. Ouyang et al. presented an approach which focused on regression models to query-focused multi-document summarization by Support Vector Regression (SVR). Another attempted is presented in [12] focused on query-oriented multi-topic based summarization as a solution of multi-topic query problem. In [13], Rasim Alguliev et al. proposed an evolutionary approach which combined the extracted sentences from source documents. In [14], S. Prabha et al. have introduced a document summarization technique named as CBSA. Context-

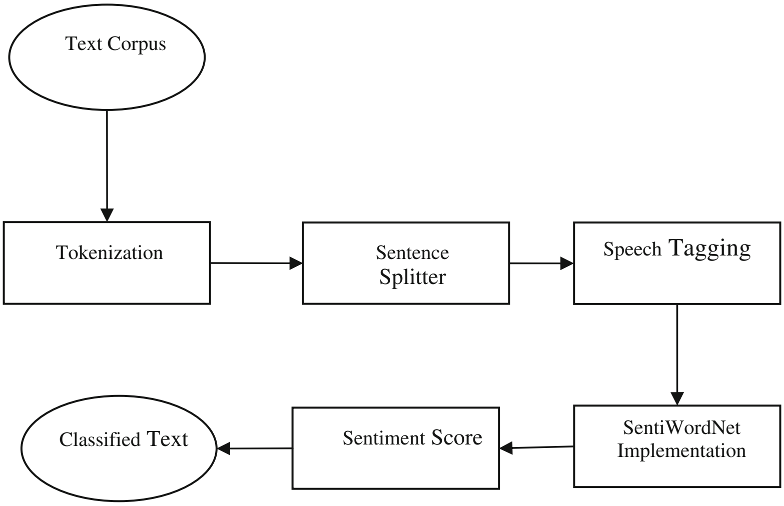


Fig. 1 Process of sentiment scoring

Based Similarity Analysis (CBSA) uses the concept of lexical association among terms which is applied to get the concept sensitive weight to terms of document. Hybrid approach using K-means and PSO gives the good result for grouping of particles in compare to PSO and K-means individually [15]. In another study, Chen Li et al. concentrate to get the better performance after resolving the issues of sentence comparison for multi-document summarization [16].

2 Sentiment Scoring

The term sentiment can be used to indicate certain emotions towards a particular object or person and it may be found in writing like product review, websites, blogs, forums, etc. [17]. Sentiment Analysis (SA) is a technique that plays an important role in the opinion mining and has a significant presence in text mining and linguistics [18]. SA can be performed in document level, sentence level and word-level. For calculating the sentiment in sentence level, SentiWordNet is used which is an open source lexical resource for opinion mining. The process of calculating sentiment score has been shown in Fig. 1.

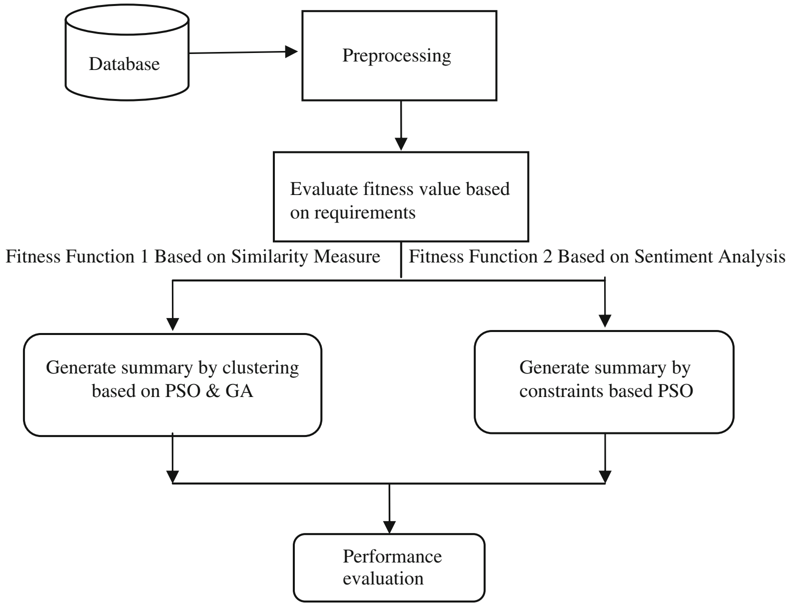


Fig. 2 System architecture for proposed technique

3 System Architecture

The architecture of the proposed technique has been given in Fig. 2. In the first step, the source document is preprocessed and then fitness value which would be used by the PSO has been calculated by sentiment score. In the next step, two sets of summary have been generated namely summary by clustering based PSO and GA and constraints based PSO. Finally, performances of both algorithms have been evaluated.

4 Methodology

4.1 Preprocessing

Preprocessing is performed on collected documents from database so that further analysis can be performed. In preprocessing, documents are represented by a group of sentences. The sentences are a collection of terms after removing the special character and stop words [1].

4.2 Evaluation of Fitness Value and Cluster Selection

After representing source documents in terms of sentences, clustering technique is applied to have group of documents of similar type. To select the best particle in the PSO algorithm, a fitness function is required. In the proposed approach, sentiment score is used to evaluate the best particle. Sentiment score is measured by SentiWordNet given by Eq. 1.

$$\mathrm{SetiScore}(\mathrm{s_i}) = \sum_{k=1}^{n} \mathrm{senti}(\mathrm{t_k}) \tag{1}$$

where $\mathrm{s_i}$ is ith sentence and $\mathrm{t_k}$ is kth term of ith sentence.

Good summarization techniques should fulfill the following criteria [19]

- Coverage: It specifies that the summary should cover almost all valuable aspects of the source document. Summary should have maximum coverage and minimum loss of information.
- Diversity: It focuses on the redundancy. A rich summary is that which can minimize the sentences with similar meaning. In summarization, process diversity minimizes the redundancy fruitfully among sentences.
- Balance: The main objective of a balanced summary is to focus on the different important aspects of the source document whereas an unbalanced summary may indicate the wrong representations of general ideas of the source document.

After getting all the clusters the following constraints like coverage, diversity and length of summary are checked. The diversity constraint is given by Eq. 2.

$$\sum_{\mathrm{s_i},\mathrm{s_j}\in \mathrm{S}} \mathrm{sim}\left(\mathrm{s_i}, \mathrm{s_j}\right) \le \Theta_{\mathrm{driver}} \tag{2}$$

where $\mathrm{i} \ne \mathrm{j}$ and Θ_{driver} is the threshold of diversity (the diversity constraint). The higher value of Θ_{driver} means low level of diversity.

$$\Theta_{count} = sim\left(o, o^{s}\right). \sum_{i=1}^{n} sim(o, s)x_i \tag{3}$$

The constraints of content coverage may be written as

$$\mathrm{sim}(\mathrm{S}, \mathrm{D}) \ge \Theta_{\mathrm{count}} \tag{4}$$

where θ_{count} is content coverage (the content coverage constraint) and o and $\mathrm{o^s}$ indicate the mean of original document and summary respectively. The higher value of θ_{count} corresponds to higher level of similarity between generated summary and source document.

The length constraint of the summary is represented by Eqs. 5 and 6 gives the integrity constraint

$$\sum_{i=1}^{n} l_i x_i \leq L \tag{5}$$

$$x_i \in \{0, 1\} \tag{6}$$

where L and l_i indicate the length of the summary and sentence s_i respectively.

4.3 Clustering Algorithm Based on PSO

The PSO based clustering algorithm for text summarization is proposed in [20]. The number of clusters k in this approach is estimated by Eq. 7.

$$k = \mathrm{n}\frac{|\mathrm{d}|}{\sum_{\mathrm{i=1}}^{\mathrm{n}} |\mathrm{S_i}|} = \mathrm{n}\frac{\left|\mathrm{U}_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{S_i}\right|}{\sum_{i=1}^{n} |S_i|} \tag{7}$$

where $|d|$ is the number of terms in document d and n is number of sentences in d. Once cluster number is determined, automatic population partitioning (APP) is implemented for sentence clustering. The PSO has been applied in APP model and the fitness value is calculated by the cluster similarity by using Eq. 8.

$$\mathrm{AverageSimi(i)} = \frac{\sum_{\mathrm{j=1}}^{\mathrm{k}} \mathrm{ClusterSim_j}}{\sum_{\mathrm{j=1}}^{\mathrm{k}} \mathrm{ClusterSize_j}} \tag{8}$$

where

$$\mathrm{ClusterSim_j} = \sum_{\mathrm{d_{jm}} \in \mathrm{C_{ij}}} \sum_{\mathrm{d_{jn}} \neq \mathrm{d_{jm}}} \mathrm{Sim}\left(\mathrm{d_{jm}}, \mathrm{d_{jn}}\right) \tag{9}$$

Here, $\mathrm{ClusterSim_j}$ indicates total similarity of each sentence within the jth cluster and $\mathrm{Sim}\left(\mathrm{d_{jm}}, \mathrm{d_{jn}}\right)$ represents the cosine similarity between $\mathrm{s_{jm}}$, $\mathrm{s_{jn}}$ sentences.

$\mathrm{ClusterSize_j}$ given by Eq. 10 indicates the number of similar sentences in the jth cluster and $|c_{ij}|$ represents the number of sentences in the jth cluster in the ith chromosome. So Eq. 8 shows if the fitness value is high then best chromosome will be chosen for solution to clustering problem.

$$\mathrm{ClusterSize_j} = \frac{\left|\mathrm{C_{ij}}\right| \times \left(\left|\mathrm{C_{ij}}\right| - 1\right)}{2} \tag{10}$$

4.4 Proposed Algorithm

The proposed algorithm follows basic concepts of PSO and fitness function is replaced with sentiment score. The particle velocity and position are updated by using this equation

$$\mathrm{Vid}(t+1) = \mathrm{Vid}(t) + c1 * r1d * (\mathrm{Pid} - \mathrm{Xid}) + c2 * r2d * (\mathrm{Pgd} - \mathrm{Xid}) \quad (11)$$

$$\mathrm{Xid}(t+1) = \mathrm{Xid}(t) + \mathrm{Vid}(t+1) \quad (12)$$

To calculate the relative distance (D) between fitness value f and best local fitness value $_{P}$*Best, i.e., indicates the search space*

$$D = \frac{(\mathrm{fmax} - \mathrm{f})}{(f_{max} - f_{min})} \quad (13)$$

where f_{max}, f_{min} indicate the maximum and minimum value of fitness in the current cluster. From Eq. 13, two cases have been depicted.
Case 1: If $f_{max} \approx f$ then the probability is high to perform a local search by particle $s_{i.}$
Case 2: If f_{max} and f differ more than particle s_i performs the global search.
The proposed algorithm can be summarized in the following steps:

1. Construct Matrix *A* from unstructured text document.
2. Define max-cluster as any arbitrary number *k* by Eq. 7.
3. Initialize each particle with the cluster centroid $O_i = \{O_1, O_2, \ldots, O_k\}\ \forall\ k > 1$
4. Initialize the seed particle with random position and velocity.
5. Loop
6. For all data point does

    6.1 Assign each data point to the nearest cluster centroids.

    $$d(O_i, x_i) < d\left(O_i, x_j\right) \quad i \neq j, i = 1, 2, \ldots, n. \text{ where } d(O_i, x_i) = \sqrt{}(O_i - x_i)^2$$

    6.2 Recalculate each cluster center to be equal to the average sentiment of all vector points within that cluster.
    $O_i \leftarrow \frac{1}{n}\sum_{j=1}^{n} sentix_j \quad i = 1, 2, \ldots, n$, where $\mathrm{senti}(x_i)$ is calculated by Eq. 1.
    if O_i == $\mathrm{senti}(x_i)$ then centroid = vector point(x_i)
    else centroid=nearest vector point

    6.3 *pBest* $\leftarrow$ Oi

    6.4 Calculate sentiment score (unsigned) of a sentence s_i as fitness f using SentiWordNet.
    $f \leftarrow \mathrm{senti}(s_i)$

    6.5 Then by Eq. 13, the search place is selected and that will reduce the complexity of the model.

Table 1 Average precision, recall and F-Measure using ROUGE-1

Dataset	Precision	Recall	F-Measure
Dataset 1	42.9	43.67	43.28
Dataset 2	45.7	42.97	44.29
Dataset 3	42.1	50.3	45.84
Dataset 4	47.02	43.9	45.41
Dataset 5	39.9	42.4	41.11
Average	43.52	**44.65**	43.99

6.6 *if* $f(O_i)<pBest$ *then*
   $pBest \leftarrow f(O_i)$
6.7 *if* $pBest<gBest$ *then*
   $gBest \leftarrow pBest$
6.8 Update particle velocity and position using Eqs. 11 and 12.

7. Save the all clusters centroids with highest fitness value.
8. End loop.
9. Until (maximum iteration > maxIteration or noChange(gBest)).
10. Return all the k clusters with fitness value in ascending order.
11. Final cluster is selected by Eqs. 2 and 3.
12. The sentences (data points) are arranged in descending order of sentiment score.
13. Topmost sentences (considering the length constraints) are included to final summary by Eq. 5.
14. Output the final summary.

5 Result and Analysis

The experiment of this model is performed on the five different datasets collected from several websites and the optimized summary has been generated. To evaluate the performance of the model with user-generated summary, F-measure is used. F-measure uses the generic metric and ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation use to compare the quality of summary. Here, ROUGE-N and N = 1 and 2, i.e., ROUGE-1 and ROUGE-2 are best for single document summarization (Tables 1 and 2).

The average precision, recall, and F-measure of swam model are 47.741, 43.028, and 44.669 for ROUGE-1, respectively. By using ROUGE-2 average precision, recall and F-measure of swam model are 21.622, 18.828, and 19.776 respectively (Fig 3).

Table 2 Average precision, recall and F-Measure using ROUGE-2

Dataset	Precision	Recall	F-Measure
Dataset 1	21	19.3	20.11
Dataset 2	23.9	20.1	21.84
Dataset 3	21.1	19	19.1
Dataset 4	20.01	16.7	18.48
Dataset 5	21.03	17.5	19.09
Average	21.546	18.52	**19.906**

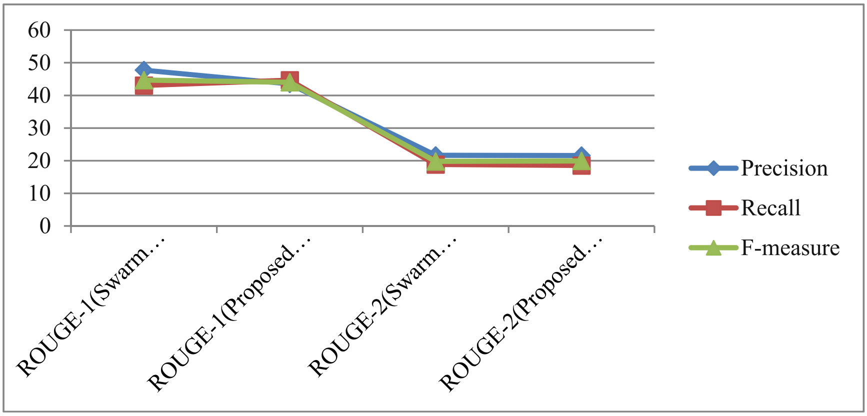


Fig. 3 Evaluation graph of the proposed method

6 Conclusion

The proposed model is an extractive based summarization technique which works based on the concepts of PSO, constraints of summary, and sentiment analysis. The basic concept is to select the most deserving cluster after satisfying diversity and coverage constraints arranging the sentences within the cluster in respect to sentiment score in decreasing order. Finally, choose the topmost sentences following the length constraints. Our experimental result says that recall and f-measure for ROUGE-1 and ROUGE-2 have been improved respectively. In future, we can improve the rest of the parameter.

References

1. Raj Kumar, V.S., Danu, R., Shanmugapriya, S., Vinod, R.: Text summary using modified particle swarm optimization algorithm. Asian J. Appl. Sci. **04**(06) (2016). (ISSN: 2321–0893)
2. Mani, I., Maybury, M.T. (eds.): Advances in Automated Text Summarization. The MIT Press, Cambridge, MA (1999)

3. Hovy, E., Lin, C.Y.: Automatic text summarization in SUMMARIST. In: Proceedings of the AC“97/EACL“97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, pp. 18–24 (1997)
4. Mihalcea, R., Tarau, P.: An algorithm for language independent single and multiple document summarization. In: Proceedings of the International Joint Conference on Natural Language Processing, Korea (2005)
5. Hirao, T., Sasaki, Y., Isozaki, H.: An extrinsic evaluation for question-biased text summarization on qa tasks. In: Proceedings of NAACL workshop on Automatic Summarization (2001)
6. DeJong, G.F.: Skimming stories in real time: an experiment in integrated understanding. Doctoral Dissertation. Computer Science Department, Yale University (1979)
7. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks, Australia, vol. 95, pp. 1942–1948
8. Alguliyev, R.M., Aliguliyev, R.M., Isazade, N.R.: An unsupervised approach to Generating generic summaries of documents. Appl. Soft Comput. **34**, 236–250 (2015)
9. Xi, M., Sun, J., Xu, W.: An improved quantum-behaved particle swarm optimization algorithm with weighted mean best position. Appl. Math. Comput. **205**, 751–759 (2008)
10. Babar, S.A., Patil, P.D.: Improving performance of text summarization. In: International Conference on Information and Communication Technologies (ICICT) (2014)
11. Ouyang, Y., Li, W., Li, S., Lu, Q.: Applying regression models to query-focused multi-document summarization. Inf. Process. Manage. **47**(2), 227–237 (2011)
12. Tang, J., Yao, L., Chen, D.: Multi-topic based query-oriented summarization. In: Proceedings of the 9th SIAM International Conference on Data Mining, Nevada, USA, pp. 1148–1159 (2009)
13. Alguliev, R., Aliguliyev, R.: Evolutionary algorithm for extractive text summarization. Intell. Inf. Manag. **1**, 128–138 (2009)
14. Prabha, S., Duraiswamy, K., Priyanga, B.: Context-based similarity analysis for document summarization. Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) **3**(4) (2014)
15. Sarkar, S., Roy, A., Purkayastha, B.S.: A comparative analysis of particle swarm optimization and K-means algorithm for text clustering using Nepali WordNet. Int. J. Natural Lang. Comput. (IJNLC) **3**(3) (2014)
16. Li, C., Liu, Y., Liu, F., Zhao, L., Weng, F.: Improving multi-documents summarization by sentence compression based on expanded constituent parse trees. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 691–701, 25–29 Oct 2014
17. Richmond, W.K.: Teachers and Machines: an Introduction to the Theory and Practice of Programmed Learning. Collins (1965)
18. Shaikh, M.A., Prendinger, H., Mitsuru, I.: Assessing sentiment of text by semantic dependency and contextual valence analysis. In: Presented at the Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction, Lisbon, Portugal (2007)
19. Li, L., Zhou, K., Xue, G.-R., Zha, H., Yu, Y.: Enhancing diversity, coverage and balance for summarization through structure learning. In: Proceedings of the 18th international conference on World Wide Web, Madrid, Spain, pp. 71–80 (2009)
20. Lee, J.S, Hahm, H.H., Park, S.S.: Less-redundant text summarization using ensemble clustering algorithm based on GA and PSO. In: Wseas Transactions on Computers, vol. 16 (2017)

Face Recognition Using Eigenfaces

G. Md. Zafaruddin and H. S. Fadewar

Abstract In this paper, we propose a PCA-based face recognition system implemented using the concept of neural networks. This system has three stages, viz. pre processing, PCA and face recognition. The first stage, preprocessing performs head orientation and normalization. The aspects that matter for the identification process are ploughed out using Principal Component Analysis (PCA). Using the initial set of facial images, we calculate the corresponding eigenfaces. Every new face is presented into the face space and is characterized by weighted-sum of corresponding eigenfaces that is used to recognize a face. To implement this face recognition system, we have created a database of faces with the help of neural networks and we have built one separate network per person. We obtain a descriptor by projecting a face as input on the eigenface space, then that descriptor is fed as input to the pre-trained network of each object. We select and report that which has the max output provided it passes the threshold already defined for the recognition system. Testing of the algorithm is done on ORL Database.

Keywords PCA (Principal component Analysis) · Neural network · Eigenface Eigenvector

1 Introduction

Faces are used by us humans to recognize people; same can be done with the help of computers automatically. Earlier, simple mathematical models were used for face recognition but in this century, face recognition has been established as a science of its own. With the advent and use of engineering in this area, face recognition systems

G. Md. Zafaruddin
Millennium Institute of Management, Aurangabad, India
e-mail: zafar.mohammed@rediffmail.com

H. S. Fadewar (✉)
Department of Computational Sciences, SRTMU, Nanded, India
e-mail: fadewar_hsf@yahoo.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_87

have been one of the major attractions in computers. Face recognition systems are used in both ways, viz. for face verification as well as face identification. There has been an extensive study of face recognition in the past 25 years with relevance to various domains like two-dimensional, three-dimensional as well as videos which has contributed immensely in the areas of research and development of the same. Still, some factors like varying posture, face expression, poor lightning, person's age, any unwanted obstacle, etc., degrade the performance of a face recognition system [1].

The field of FRS provides with a host of prospects that can be exploited for advancement of research. Thus, our study tries to address the reasons that depreciate the performance of FRS and also explores better techniques. The proposed system tries to remove various factors that affect the performance of FRS like varied postures, poor lighting, and different expression, etc. [2] FRS is simply pattern recognition for faces that can be used to differentiate as an unknown and known face. Because face is a dynamic biometric it has huge number of problems to solve. The developers and researchers in the field of image procession, human–computer interaction and AI have come out with many alternative remedies and solutions to reduce the problematic factors and make the FRS more robust and accurate.

In broader sense, the approach to face recognition may be differentiated as either being feature-based or holistic-based. The set of features used in both approached is basically different. In the holistic approach, recognition process is carried out on the basis of globally extracted features from a face, whereas the feature-based approach uses the local features. The holistic-based techniques signify the optimum variance of data in pixels of face images used to identify a subject and features of face like mouth, nose and eyes are used in feature-based approach for the identification process [3].

2 FR Using Eigenfaces

The main inspiration for creating eigenfaces was face recognition. So, eigenfaces are having an edge over different available techniques for the same with respect to the effectiveness and pace of the face recognition systems. Basically, the eigenface approach follows dimension reduction method, hence an FRS can easily embody data of a number of subject with a very small-sized data. There is no significant impact on the performance of the FRS due to reduction in image size but FRS fails significantly with variation in the probe versus seen images [4].

For the recognition process, the images seen by the FRS are stored as collection of weights that describe the eigenfaces for the said image [3]. As a new input faces is given to the FRS for identification, the corresponding weights of that image is calculated by representing the image as a collection of eigenfaces resulting in weights of the input image which is to be probed. Then, the weights are compared against all the stored weights in the database with an aim to identify the match which is closest.

Fig. 1 Facial sets

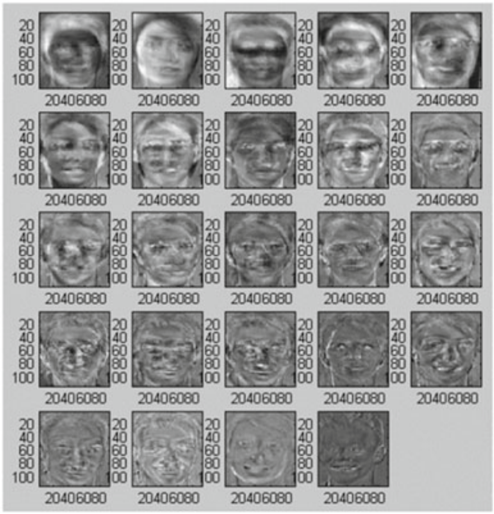


Fig. 2 Matching eigenfaces

This nearest-neighbour is a very simple technique to find the Euclidean Distance between two vectors and the min is used to classify as the closest [4].

Facial sets, their average face, eigenface of each facial image and the matching eigenvalues are represented in the figures displayed beneath. Every eigenface deviates from uniform grey where facial feature differs in the training set. Eigenfaces are nothing but a type of map of variation among faces [5, 6] (Figs. 1, 2, 3 and 4).

Consider a facial image X to be a 2D N $\times$ N matrix of 8bits. Consider the image to be a vector having a dimension value of N2. For (example), a common image having dimension of 256 $\times$ 256 can become a vector of 65536. Then, an image group is

Fig. 3 Reconstruction of first image with eigenface

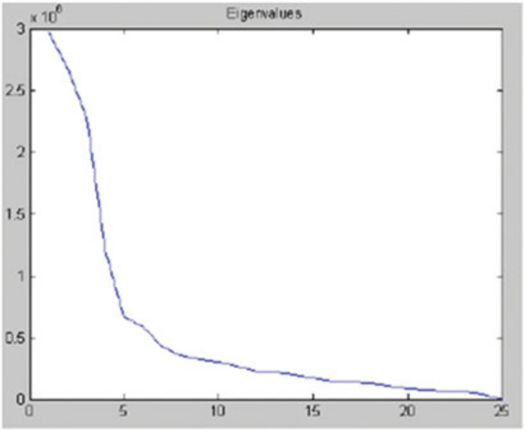


Fig. 4 Corresponding eigenvalues

mapped as a group of points. As facial images are same on the whole configuration are not distributed in random for the image space and are described through very low-dimensional-subspace. Basic aim of PCA is finding vectors which are important for distributing facial images in the whole image space [2].

Now, those vectors are used in defining subspace of facial images that is being called as a face space. Every vector is having a length of N that describes an N $\times$ N facial image, and represents linear combination of the original facial image. These image vectors are eigenvectors of covariance matrix which corresponds to the original facial image, having face like representation (appearance), hence we call them eigenfaces. For (example), eigenfaces shown in the Fig. 2.

3 Neural Network Simulation

For performing composite functions in different areas of applications like speech control, vision control, identification and classification and pattern recognition neural networks are being trained [7]. We have built a separate neural network for every person's image present in the facial database. Once eigenfaces are obtained, the corresponding calculation for obtaining feature vectors for the facial images in the database is done and is provided as input for training each neural network.

The training algorithm uses the face feature vectors of the same individual person which are used to train an individual's neural network and also other neural networks. Whenever an input image is given for facial recognition process, then the corresponding feature vectors are computed using already calculated eigenfaces and the new descriptors of input image are obtained [5]. These descriptors are given as input to all neural networks and those neural networks are replicated with the descriptors. Then, comparison of output given by neural network is done. In case, max output value overcomes the already defined threshold value, the input facial image is considered to be of that person having max output.

4 Summary of Eigenface Based Face Recognition

Summary of the eigenface-based face recognition approach is presented below:

- Create a facial image database of known-persons.
- Decide a training set with M number of images corresponding to every individual having disparity in facial expressions and lighting conditions.
- Calculate MxM matrix (L) and the corresponding eigenvectors and its eigenvalues. Select M' eigenvectors having the highest corresponding eigenvalues.
- Merge normalized images training set that produces M' eigenfaces and save the corresponding values.
- Compute and save a feature-vector for each individual in the database.
- Build a neural network of every individual present in the facial image database.

5 Experimental Results

We have used ORL face image database to test our method. ORL face image database is having multiple facial images of every individual under varying conditions. Next, we describe various details for ORL database and the related performance of the proposed FRS. Here, we have used a separate neural network for each individual in the database.

We have used ORL facial database for testing the proposed method with regular occurrence of head position disparity. All facial images have been captured against

Fig. 5 ORL face database

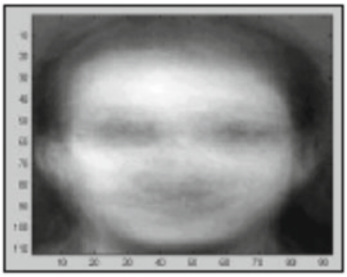


Fig. 6 Mean face for ORL face database

a dark homogeneous background. For 40 different persons, 10 differing images are taken with respect to

- diverse timing,
- altering lighting conditions,
- varying face expressions, viz. open eyes/closed eyes and smile/no smile,
- varying face detail, viz. with/without specs,
- head pose, viz. tilt and rotation.

Figure 5 given below represents ORL database with whole set of 40 persons with 10 different images of each individual. As total neural networks are equal to total number of individuals in the facial image database, 40 neural networks, viz. 1 for every individual have been built. In the given 10 face images, initial 4 have been taken to train neural networks, and testing of those neural networks is done. After testing, the features of those networks are updated to get min squared-error-function. Now, the trained networks are to be utilized for facial recognition process.

Mean face for entire ORL face database, their corresponding eigen values and the ORL facial image database's top-30 eigenfaces represented in Figs. 6, 7 and 8 in that order.

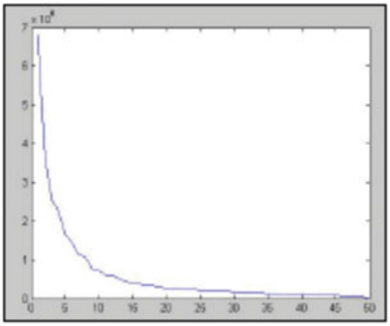


Fig. 7 The corresponding eigenvalues

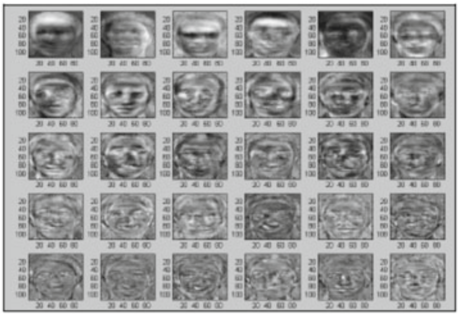


Fig. 8 The ORL eigenfaces

The overall efficiency of face recognition is enhanced with respect to number of facial images utilized to train the neural networks. The table given below represents the recognition rates with respect to varying number of images utilized to train the networks (Table 1).

Facial recognition rate is affected by the neurons present in the hidden layer and eigenfaces which have been used for describing a face, hence we have conducted the tests with varying number of both parameters and the corresponding results are given in Table 2 presented below. Contrast of the facial image is enhanced considerably with histogram equalization by converting corresponding values in the face image. Here, we represent recognition rates in the Table 2 given below without/with histogram equalization.

Table 1 Recognition rates with respect to varying number of images to train and test networks

Images in training (per individual)	Images in testing (per individual)	Eigenfaces	Hist. Equ.	Recognition rate (%)
1	9	10	Done	8.8
1	9	,,	Not done	9.7
2	8	20	Done	25.0
2	8	,,	Not done	26.0
3	7	25	Done	51.7
3	7	,,	Not done	51.7
4	6	30	Done	75.0
4	6	,,	Not done	76.2
5	5	50	Done	58.5
5	5	,,	Not done	90.0
6	4	60	Done	90.0
6	4	,,	Not done	92.5
7	3	100	Done	89.1
7	3	,,	Not done	91.6
8	2	,,	Done	88.8
8	2	,,	Not done	91.2
9	1	,,	Done	87.5
9	1	,,	Not done	90.0

Table 2 Representing recognition rate with varying neurons and Eegenfaces (Histogram Equalization done/five Images used in training and testing)

Eigenfaces	Hidden layer consisting	Recog. rate (%)
40	5 Neurons	40
	10 Neurons	58.3
	15 Neurons	72
	20 Neurons	75.0
50	5 Neurons	54.8
	10 Neurons	73
	15 Neurons	87.5
	20 Neurons	90.8
60	5 Neurons	48.8
	10 Neurons	81.3
	15 Neurons	89.5
	20 Neurons	92.8
70	5 Neurons	50
	10 Neurons	86.3
	15 Neurons	92.3
	20 Neurons	93.3

Table 3 Comparative efficiency analysis of ORL database

Approach	Recog. rate (%)
LDA	80
Bayesian PCA	93
Proposed method	93

6 Conclusion

Proposed method is applied on ORL Database. We have implemented varying number of neurons in the hidden layer and varying number of eigenfaces using neural networks. We find here that for the entire database with five images of every individual to train the network, hidden layer consisting 15 neurons and 50 eigenface values for entire face database are adequate for acceptable recognition rates of about 93%.

Eigenface approach is quite susceptible to head position disparity. Facial image variance happens for those face images which have significant head position disparity. We represent comparative analysis of the proposed method and the already established approaches in Table 3 given below.

References

1. Rathi, R., Chaudhary, M., Chandra, B.: An application of face recognition system using image processing and neural networks. Int. J. Comput. Technol. Appl. **3**(1), Jan-Feb 2012
2. Jafri, R., Arabnia, H.R.: A survey of face recognition techniques. J. Inf. Process. Syst. **5**(2), June 2009
3. Sung, K., Poggio, T.: Example-based Learning for View-based Human Face Detection. A.I. Memo 1521, CBCL Paper 112, MIT, Dec 1994
4. Sellahewa, H., Jassim, S.A.: Image quality-based adaptive face recognition. In: IEEE Transactions on Instrumentation and Measurement and Measurement, pp. 805–813. IEEE (2010)
5. Patil, M., Iyer, B., Arya, R.: Performance evaluation of PCA and ICA algorithm for facial expression recognition application. In: Proceedings of Fifth International Conference on Soft Computing for Problem Solving, pp. 965–976 (2016)
6. Shermina, J.: Face Recognition System using Multi Linear Principal Component Analysis and Locality Preserving Projection. In: IEEE GCC Conference and Exhibition, 19–22 Feb, pp. 283–286. Stirling, UK (2011)
7. Sirovich, L., Kirby, M.: Low-dimensional procedure for the characterization of human faces. J. Opt. Soc. Am. **4**, 519–524 (1987)
8. Lone, M.A., Zakariya, S.M., Ali, R.: Automatic face recognition system by combining four individual algorithms. In: International Conference on Computational Intelligence and Communication Systems IEEE, pp. 222–226 (2011)
9. Turk, M., Pentland, A.: Eigenfaces for Recognition. J. Cogn. Neurosci. **3**, 71–86 (1991)
10. Dimitri: Eigenface-Based Facial Recognition (2013). http://openbio.sourceforge.net/resources/eigenfaces/eigenfaceshtml/facesOptions.html. Accessed 13 Feb 2003
11. Rowley, H.A., Kanade, T.: Neural network-based face detection. IEEE Trans. Pattern Anal. Mach. Intell. **20**(1), 23–30 (1998)
12. Fraud, R., et al.: A fast and accurate face detector based on neural networks. IEEE Trans. Pattern Anal. Mach. Intell. **23**(1), 42–53 (2001)

13. Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D.: Face recognition: a convolutional neural network approach. In: IEEE Transactions on Neural Networks, Special Issue on Neural Networks and Pattern Recognition, pp. 1–24 (1997)
14. Galbally, J., McCool, C., Fierrez, J., Marcel, S., Ortega-Garcia, J.: On the vulnerability of face verification systems to hill-climbing attacks. Pattern Recogn. **43**(3), 1027–1038 (2010)

Multi-focal Image Fusion with Convolutional Sparse Representation and Stationary Wavelet Transform

Gandhali A. Pawar and Sujata Kadam

Abstract This paper illustrates a completely unique technique of multi-focus image fusion involving Stationary Wavelet Transform (SWT) and Convolutional Sparse Representation (CSR). Sparse-based fusion strategies do not retain information representation and cannot tolerate minor mistakes in registration. The SWT method does not have these issues. Multi-focus image fusion is the fusion of different parts of digital images, representing the common scene, in order to produce an image with everything in Focus, i.e., without the blur effect. Camera processors cannot fuse images by themselves. Thus, experts have to employ image editing methods to obtain clear photographs. The scheme stated in this paper uses SWT to distinguish focus levels accurately. The results suggest that the strategy is successful in ways comparable in terms of visual quality and clarity.

Keywords Image fusion · Convolutional sparse representation · Stationary wavelet transform · Minute loss prevention · Shift tolerance

1 Introduction

1.1 Introduction

Multi-focus image fusion involves fusing the most focused parts of several images of the common scene, to extract a combined image of the scene where all parts are in focus. This can be done manually, wherein the photographer first takes several photographs of the same scene using different camera settings in order to get different focus levels for each part of the scene. Once this is done, the photographer uses an image editing software to get a new image where the entire scene is in focus. One

G. A. Pawar (✉) · S. Kadam
RAIT, Nerul, Navi Mumbai, India
e-mail: gandhalivjadhav@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_88

of the issues with this technique is the fact that the photographer has to retake the photographs should he notice that the initial photographs are not working out. There are many automatic ways of performing multi-focus image fusion. One of these ways uses wavelet functions which are used to determine what is in focus and what is not. The end results are usually accurate, however, the technique is advanced and is therefore difficult to implement. Also, the input images have to align perfectly in order for the technique to work. This is difficult for the photographer to achieve, and if it is not achieved then manual post-processing of the photographs is needed.

2 Related Work and Motivation

2.1 Related Work

Stationary Wavelet Transform in combination with CSR promises better Multi-focal image fusion results. A number of image fusion methods are proposed. Generally, these strategies are often classified as domain spatial and domain transform [1]. Spatial process addresses the merging problem through windowing process [2–5] or picture element gradient values [6–9]. This particular class is appropriate for merging tasks wherein input pictures are procured through a common sensor, as in varied focus merging [2, 4–6, 9] and multiexposure fusion [3, 7, 8]. Processes of transform domain combine the details containing images after remodeling to the next level. The combined picture is created by applying restructuring procedures of combined coefficients. General remodels researched within the field of picture merging involve discrete wavelet transform (DWT) [10], Laplacian pyramid (LP) [11], etc. Because the picture type methods sync with the natural operation of the human eyes mechanism, strategies of this field have seen quite sensible fusion of multimodal picture [12, 13], during which the images are acquired by numerous varieties of imaging devices. Another benefit of remodel domain is that it can handle the merging of grainy pictures while combining noise removal steps with the merging process [14, 15]. During this work, we tend to primarily study remodel domain methods.

2.2 Stationary Wavelet Transform (SWT)

Stationary Wavelet Transform is a translation invariant alteration of the Discrete Wavelet Transform that does not decimate coefficients at every transformation level. Stationary Wavelet Transform is also called as wavelet transform that is undecimated.

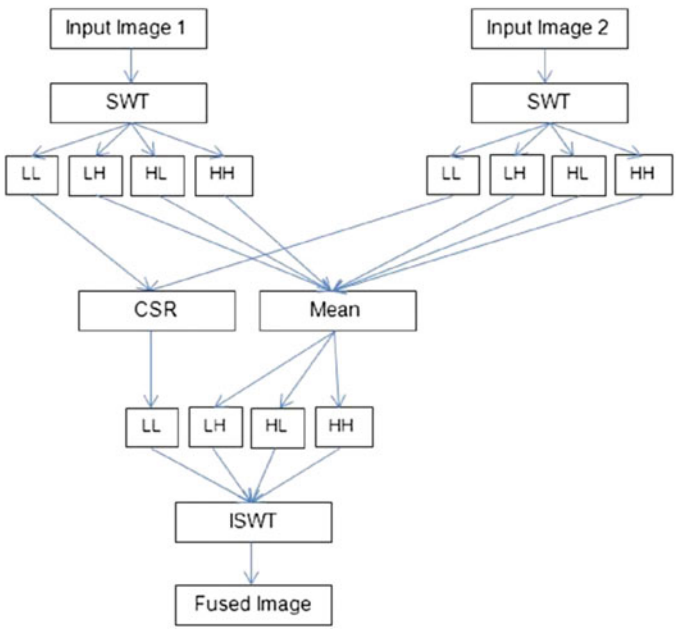
3 Proposed Fusion Scheme

3.1 Proposed Fusion Scheme

Resolution retention is one of the most important concerns in the image fusion process. CSR is unique and focused on maximizing Fusion Quality thus improving coverage of the complete image.

The two multi-focal input images are first processed for Stationary Wavelet Transform (SWT). SWT provides us with four subbands, namely LL, LH, HL, and HH. CSR is employed on LL bands to get the fused output. Mean is employed on the other bands to get the respective output subbands. Inverse SWT is employed to get the final output fused image.

Say there are N previously registered input images denoted as S_n, $n \in \{1, \ldots, N\}$, and say a set of lookup-filters d_m, $m \in \{1, \ldots, M\}$ are trained with the lookup way of learning [10]. The proposed SWT-based image fusion framework would have the following steps.



3.2 Stationary Wavelet Transform

SWT was essentially developed to get over the issues of transform fixation of the DWT or Discrete wavelet transform. Transform fixation is reached while eliminating down values along with the up values occurring in the wavelet transform and up mapping the filter values with a constant of $2^{(j-1)}$ on jth layer. It is an implicitly recurring system as the output of each level in SWT has the same No. of samples as is the count seen in the incoming values going towards division of all the layers.

3.3 Two-Scale Image Decomposition

Every input image S_n is first transformed into a base level S_n^b and a minute level S_n^d. The base level is got by the Tikhonov regularization optimization method which involves fast Fourier transform (FFT). Having S_n^b, the minute level can be acquired by subtraction: $S_n^d = S_n - S_n^b$. This way of two-scale decomposition is well under research that has been ongoing in various image fusion topics such as [3].

3.4 Fusion of Minute Levels

For every minute level S_n^d, its sparse coefficient maps $D_{\{n,p\}}$, p in [1. ... P] are retrieved by considering the CSR model with the method in [12].

Let $D_{n,1:M}(u, v)$ denote the components of $D_{n,m}$ at the position (u, v) in spatial domain. Clearly, $D_{n,1:M}(u, v)$ becomes an M dimensional vector. According to the method used in the SR-based fusion methods [14, 15], the l1-norm of $D_{n,1:M}(u, v)$ is chosen as the minute level assessment of input images. Thus, the phenomena level map $E_n(u, v)$ is retrieved by

$$E_n(u, v) = \left\| D_{n,1:M}(u, v) \right\|_1$$

In order to avoid overlapping mistakes and its effects, block mean method is applied on $E_n(u, v)$ to get the final minute level map.

The procedure is quite insensitive to overlapping mistakes with a bigger r, but few small details may be lost simultaneously. In this method, element endings in various input pictures develop unique edges making the parts not totally exact, thus a much bigger r is suggested. In this method, since detailed elements usually stay, it is better to choose a lesser value of r. In this work, r changes between 3 and 9 especially in the above two methods. Later, maximum method is applied to acquire the fusion values map:

$$D_{f,1:M}(u, v) = D_{n^*,1:M}(u, v), n^* = argmax_n E(u, v)$$

Ultimately, the fusion output of minute levels is made up.

3.5 Fusion of Base Levels

Various fusion systems for base levels are employed to various kinds of merging works. In this method, the biggest main problem involves minute level acquisition. The reason for this is that few minute values are left in the lowest levels. The maximum method relies on the phenomena layer map acquired in the minute level combination that is selected. The o/p of lowest levels is computed as follows:

$$S_f^b(\mathrm{u, v}) = \mathrm{S_n} * (\mathrm{u, v}), \mathrm{n}^* = max_n(\mathrm{E_n(u, v)})$$

Somehow, choosing this method may create visual anomaly as the input images are retrieved with different properties. Also gray levels at those points can vary significantly.

Thus, averaging method is employed to the fusion of multimodal images.

3.6 Two-Scale Image Reconstruction

Having S_f^b and S_f^d, the fused image S_f is reconstructed by $\mathrm{S_f} = S_f^b + S_f^d$.

4 Experiments

4.1 Experiments

In this work, ten buffet photos, ten infrared photos, and ten medical photos are focused upon for the purpose of test images.

The layer values and fusion scheme in our work use methods given in the papers [14, 16]. For the sake of a good and efficient distinction of CSR and SR in fusion of pictures, our proposed CSR-based method is primarily seen along the methods of work shown to make CSR method meritorious. SR-SOMP [13] and SROMP [10] are the names given to these two procedures.

Four commonly known objective metrics are chosen to assess performance, namely

- *The entropy EN*:

Entropy is mathematical level of randomness which can be used for the assessment of the surface of the Input Image. Entropy is computed as—sum(p.*log2(p)) wherein the parameter p contains the leveled statistical frequency.

- *The gradient-based metric QAB/F* [12]:

This objective metric is applied to assess Image Fusion Performance.

- *The visual information fidelity fusion VIFF* [12]:

This objective metric is applied to assess information fusion visual performance.

Every measure is selected to the suggested value as found in [14, 15]. The lookup involved in the work is having 256 elements as being trained by the K-SVD work [11] involving normal picture parts. In this work, 8×8 lookup is taken which is in line with the dimensions of the lookup element in SR-OMP and SR-SOMP. The training process focused in [15] being involved to train lookup values of 50 good normal 256×256 pictures. In the assessments, the no. of lookup values are selected, namely 128, 64, 32, and 16 for analyzing the performance of the merging work. Thus, related works are designated as CSR-128, CSR-64, CSR-32, and CSR-16. This is also referred to in the study of information on lookup training. As per the research (in picture building works) in [15], an argument of the equation selected as 1e-2 for every test (Figs. 1, 2, 3, 4, 5, 6 and Table 1).

Fig. 1 Multi-focus image: source-picture 1, source-picture 2, and CSR output

Fig. 2 Visible-infrared image: source-picture 1, source-picture 2, and CSR output

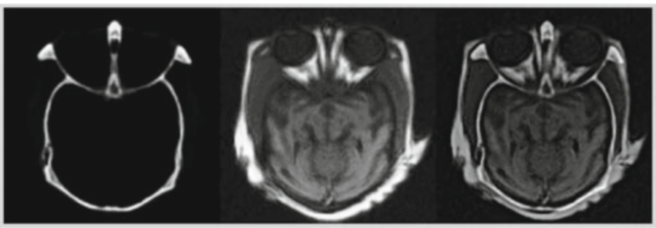


Fig. 3 Medical image: source-picture 1, source-picture 2, and CSR output

Fig. 4 Multi-focus image: source-picture 1, source-picture 2, and MCSR output

Fig. 5 Visible-infrared image: source-picture 1, source-picture 2, and MCSR output

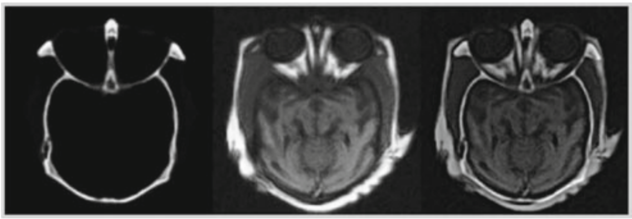


Fig. 6 Medical image: source-picture 1, source-picture 2, and MCSR output

Table 1 Comparative results

Images	Metrics	CSRF-16	CSRF-32	CSRF-64	CSRF-128	MCSRF
Multi-focus	ENT	7.3270	7.3256	7.3254	7.3255	7.3232
Multi-focus	QAB/F	0.7123	0.7115	0.7116	0.7117	0.7137
Multi-focus	VIFF	0.9283	0.9277	0.9277	0.9277	0.9387
V-infrared	EN	6.7178	6.7141	6.7144	6.7143	6.5324
V-infrared	QAB/F	0.6783	0.6824	0.6824	0.6824	0.6824
V-infrared	VIFF	0.8362	0.8420	0.8421	0.8421	0.8431
Medical	EN	6.6593	6.6623	6.6625	6.6624	6.1778
Medical	QAB/F	0.7227	0.7224	0.7225	0.7225	0.7226
Medical	VIFF	0.4699	0.4703	0.4702	0.4703	0.4704

5 Conclusion

In this work, we have successfully studied modified convolutional sparse representation (MCSR) applied to image fusion to remove limitations in CSR and Sparse-level image fusion works. A novel efficient fusion method relying on Wavelet and CSR is arrived at. Experimental outcomes in multimodal and multi-focus image fusion illustrate the merits of the suggested MCSR-based works over CSR and SR-based works. We learn that MCSR has good scope to develop into a well-worked method towards image fusion in the coming time.

The future scope for this work would be the impact of few free arguments in the proposed work, like argument of regularization of CSR scheme and the spread size of the lookup. Additionally, likewise, towards the huge developments attained in SR-based image fusion, additional meritorious schemes of fusion involving CSR holds scope and can be subsequently worked upon for improved efficiency of fusion.

References

1. Li, S., Kang, X., Fang, L., Hu, J., Yin, H.: Pixel-level image fusion: a survey of the state of the art. Inf. Fus. **33**, 100–112 (2017)
2. Goshtasby, A.: Fusion of multi-exposure images. Image Vis. Comput. **23**(6), 611–618 (2005)
3. Aslantas, V., Kurban, R.: Fusion of multi-focus images using differential evolution algorithm. Expert Syst. Appl. **37**(12), 8861–8870 (2010)
4. Li, S., Kang, X., Hu, J., Yand, B.: Image matting for fusion of multifocus images in dynamic scenes. Inf. Fus. **14**(2), 147–162 (2013)
5. Zhang, W., Cham, W.-K.: Gradient-directed multiexposure composition. IEEE Trans. Image Process. **21**(4), 2318–2323 (2012)
6. Gu, B., Li, W., Wong, J., Zhu, M., Wang, M.: Gradient field multiexposure images fusion for high dynamic range image visualization. J. Vis. Comun. Image Represent. **23**(4), 604–610 (2012)
7. Burt, P., Kolczynski, R.: Enhanced image capture through fusion. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 173–182 (1993)
8. Li, H., Manjunath, B., Mitra, S.: Multisensor image fusion using the wavelet transform. Graph. Models Image Process. **57**(3), 235–245 (1995)
9. Cao, L., Jin, L., Tao, H., Li, G., Zhuang, Z., Zhang, Y.: Multi-focus image fusion based on spatial frequency in discrete cosine transform domain. IEEE Singal Process. Lett. **22**(2), 220–224 (2015)
10. Yang, B., Li, S.: Multifocus image fusion and restoration with sparse representation. IEEE Trans. Instru. Meas. **59**(4), 884–892 (2010)
11. Zhang, Q., Guo, B.: Multifocus image fusion using the nonsubsampled contourlet transform. Signal Process. **89**(7), 1334–1346 (2009)
12. Liu, Y., Chen, X., Member, IEEE, Ward, R.K., Fellow, IEEE, Wang, Z.J., Senior Member, IEEE Image Fusion With Convolutional Sparse Representation
13. Yu, N., Qiu, T., Bi, F., Wang, A.: Image features extraction and fusion based on joint sparse representation. IEEE J-STSP **5**(5), 1074–1082 (2011)

14. Zhang, Q., Levine, M.: Robust multi-focus image fusion using multi-task sparse representation and spatial context. IEEE Trans. Image Process. **25**(5), 2045–2058 (2016)
15. Wohlberg, B.: Efficient algorithms for convolutional sparse representation. IEEE Trans. Image Process. **25**(1), 301–315 (2016)
16. Goshtasby, A., Nikolov, S.: Image fusion: advances in the state of the art. Inf. Fus. **8**(2), 114–118 (2007)

Fuzzy Deep Learning for Diabetes Detection

Tushar Deshmukh and H. S. Fadewar

Abstract The use of science for the betterment of society is the main cause for research for years. That is the reason the framework of diabetes diagnosis is always changing with new dimensions. The new and advance algorithms on the horizons are tried in hope of getting better accuracy and speed. Apart from normal algorithms researchers have tried the possible hybrid combinations. In recent times, the Convolution Neural Network (CNN) has outperformed most of the application areas of traditional prediction algorithms. Here is an attempt to use the deep convolutional neural network for diagnosis of diabetes. This work has two major contributions, first is the application of CNN for diabetes detection and second is data fuzzification in matrix form to suit needs of CNN. In the experiments, the comparison is made between classical NN and CNN for diabetes detection. Results prove that fuzzification of data significantly improves the accuracy of CNN and CNN outperforms classical NN.

Keywords Deep learning · Convolutional neural network · Fuzzy deep learning Classification · Diabetes detection

1 Introduction

The defects in the secretion of insulin, insulin action or both can cause diabetes, which is a metabolism disorder characterized by hyperglycemia [1]. The severity of disease is associated with dysfunction or damage of several organs like eyes, kidneys, nerves, and heart and blood vessels [2]. Even though there are various symptoms which signifies the disease, in many of the cases, the symptoms are either go unnoticed or no significant symptoms are shown out. So it is highly required

T. Deshmukh · H. S. Fadewar (✉)
School of Computational Sciences, SRTMU, Nanded, India
e-mail: fadewar_hsf@yahoo.com

T. Deshmukh
e-mail: gemini.tushar@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_89

that diabetes should be diagnosed in time. Various methods and techniques, and algorithms have been used for detecting diabetes.

Lofti Zedah has coined the term fuzzy logic for the first time. Fuzzy logic is used to represent the uncertainties or the approximations in reasoning [3]. Wherever one has to take decisions based on incomplete information or the information where vagueness is involved fuzzy logic can be the solution.

Artificial neural network is such parallel system which is modeled on human nervous system. It is a network of highly interconnected small processing units called neurons which learn by example [4]. The most important characteristics of ANN is their adaptability and the network can deal with incomplete, noisy data. Neural networks are considered to be the best tool for generalization.

Even though both fuzzy logic and artificial neural network are strong in their respective domains, there is always a need for a hybrid system, which is a combination of best of both worlds [5]. Fuzzy neuro systems are integration of fuzzy logic and neural network. Here, the dataset is fuzzified and then passed as input to the multilayered neural network. Then the neural network is trained on the basis of the fuzzy inputs, it then auto adjusts the weights and then produces the desired output.

The branch of machine learning that is based on different level of representation each of which corresponds to the different feature available. Whenever the data like image, text, or audio contains different level of representation, deep learning can be the best way to learn. The deep neural network simply specifies that they should have more than one hidden layer. Here each layer's training features are decided by the output of the previous layer [6]. The notable difference between normal neural network and deep neural network is that signal function. When we are using composite multiple linear function, there combine result would be another linear function. But when we are training deep neural network, we have to be very specific in choosing a separate nonlinear activation function for each of the hidden layer.

Convolutional neural network is a special kind of neural network where the complexity of multilayer neural network has been minimized. Convolutional neural network has got a good ability to abstract the data that could be used for the predictions [7]. CNN got a wide range of applications in pattern recognition and image processing.

The rest of the research work is organized as follows: Sect. 2 discusses the related work in the domain; in Sect. 3, the proposed framework is documented; whereas the Sect. 4 is about the conclusion.

2 Related Work

The related research work can be viewed as detection of diabetes using neural network, the use of deep neural network for diabetic retinopathy and deep neural network in other related area.

When the data is highly multidimensional, it becomes very difficult to store, visualize or classify such data. Hinton and Salakhutdinov, [8] in their research, suggest a multilayer encoder network which is used to transform the high dimensional data to a lower dimensional data. It is a nonlinear generalization of principal component analysis. These auto encoders with back propagation are very effective for nonlinear data reduction and that can be in both the directions between data and code space.

In 2016, Kamble and Patil [9] use deep learning approach for the prediction of diabetes. The researcher uses restricted Boltzmann machine for the classification of data that is whether the patient is diabetic or not, and once it is tagged as diabetic, then uses a decision tree to check whether it is Type 1 or Type 2 diabetes.

Convolutional neural network has been used for diabetes retinopathy by Pratt et al. [10]. The network goes on learning more features as the number of layers goes on increasing. The network could classify the data with 75% accuracy with five different classes. The research has used Keras and Theano as machine learning tool library.

Tharani S. and C. Yamini have used CNN to train the neural network for prediction of diabetes [11]. The first layer produces the feature map to forward it to the next layer where the convolution calculations are done and then find the feature value. Then input deviation is back propagated and if the feature is selected, the output layer gives result for classification.

In an attempt of automatic detection of diabetic retinopathy, Chandore Vishakha and Asati Shivam have used CNN [12]. The researcher has used three consecutive convolution layer and then a maxpool layer to increase accuracy. To train the network faster, the researchers made RELU layer after every convolutional layer. To reduce the over fitting, horizontal and vertical reflections with 50% probability are generated called data augmentation which increases the size of dataset. Also, dropout layers are used for avoiding over fitting.

Daojian Zeng, Kang Liu, and team have used the convolutional deep neural network to extract lexical and sentence level features for relation classification [13]. The researchers have shown that there is significant improvement when position features are added. The sentence level features are learned using convolutional approach.

In the paper published in 2014, Karen Simonyan and Andrew Zisserman use very deep convolutional neural network for large image classification [14]. They have used a small convolutional filter of size (3, 3). The convolutional layers are followed by multiple fully connected layers in their architecture. The very deep convolutional layer up to 19 layers proved that it benefits accuracy by large degree.

We have previously undergone a literature survey to find out, what are other techniques that have been used for diabetes prediction [15, 16]. There are a lot of work that has been done for such kind of prediction, but we can no find any related work where the researchers have used fuzzy deep learning method for diabetes detection. So in that sense, we can say it is a novel approach we are proposing.

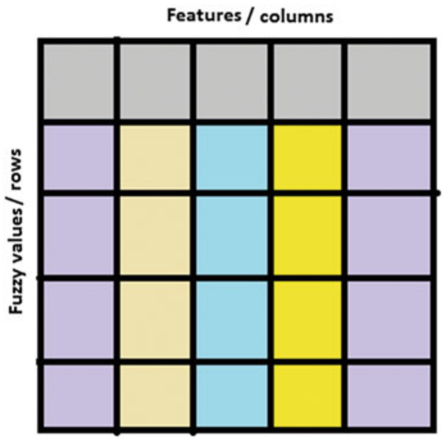


Fig. 1 Matrix form of input after fuzzification

3 Proposed Framework

In this system, convolutional neural network (CNN) is used for processing the fuzzy inputs. The unique feature of CNN is that it accepts an image like inputs and identifies unique patterns in them. In the proposed system, after fuzzification, each sample is represented as a matrix than in a row.

So, as shown in Fig. 1, each sample of original input is converted to one matrix. Matrix consists of feature as column and fuzzy value as row. So, each original sample is treated as a matrix input to CNN. This allows deeper pattern identification leading to better accuracy compared to classical Neural Networks.

3.1 Experiment and Results

Dataset used for training the network is taken from National Institute of Diabetes and Digestive and Kidney Diseases [17]. The dataset consists of nine parameters out of which eight are used as input parameter and the last one is the output which is a binary classification about diabetes or nondiabetes. There are a total of 768 samples.

In fuzzification step, each feature has defined ranges. Table 1 shows standard range of all attributes. As per the range with 10% overlapping, the fuzzification of data is done. In this experimentation, NN are configured as per the standard method. The number of neurons in first hidden layer depends on number of features, number of samples and number of outputs expected. Formula is described by Eq. 1.

$$N_{neurons} = \frac{n_{samples}}{\left(\alpha\left(n_{features} + n_{output}\right)\right)}, \text{ where } 2 \le \alpha \le 10 \tag{1}$$

Table 1 Analysis of the proposed technique

S. no	Name of attribute	Term applied		Range	Comment
1	No of times pregnant	Nulli para		No pregnancy	Not fuzzified
		Primi-para		First time pregnant	
		Multi-para		More than one pregnancy	
		Multi-para	Low	1, 2	
			Medium	3, 4, 5	
			High	>6	
2	Plasma glucose concentration a 2 h in an oral glucose tolerance test	Normal		Less than 140 mg/dl	Oral glucose test
		Prediabetes or impaired glucose tolerance		140–190 mg/dl	
		Diabetes		≥200	
		After 1 h		>140 (but if reading is >190 then without further test diabetes)	OGTT in pregnancy
		After 2 h		<155 mg/dl	
		After 3 h		<140 mg/dl	
3	Diastolic blood pressure (mm Hg)	Low		40–60	
		Ideal		60–80	
		Pre-high		80–90	
		High		90–100	
4	2 h serum insulin (mu U/ml)	Fasting		<25 mlu/L	
		30 min after glucose admin		30–230	
		1 h after glucose admin		18–276	
		2 h after glucose admin		16–166	
5	Body mass index (weight in kg/(height in m)^2)	Underweight		<18.5	
		Normal/healthy		18.5–24.9	
		Overweight		25–29.9	
		Obese		30–39.9	
6	Diabetes pedigree function	Low		<40%	
		Medium		40–80%	
		High		>80	
7	Age	Young		16–25	
		Lower middle		26–35	
		Middle age		36–45	
		Senior		>46	

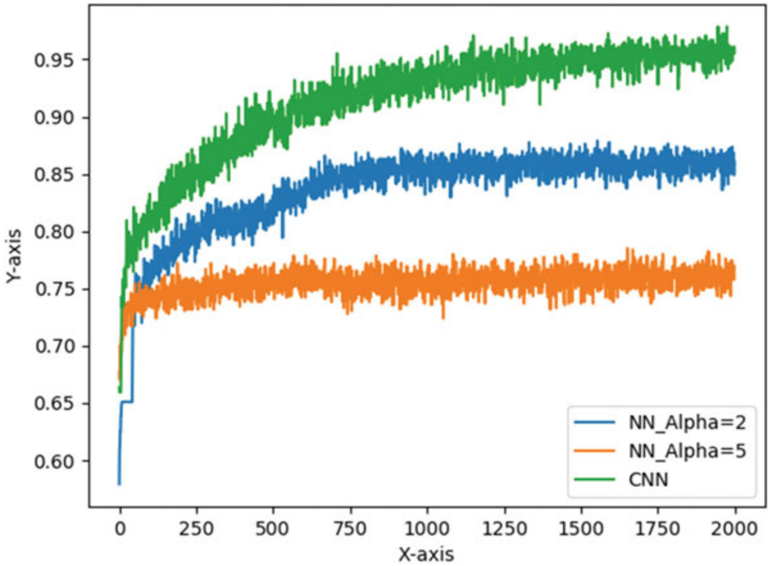


Fig. 2 Comparison of accuracy of NN ("α" = 2), NN ("α" = 5) and CNN

In this experiment, "α" has value 2 and 5. In CNN, configuration standard size of pooling matrix (2, 2) is used. Also, the convolution matrix is of size (3, 3). Number of epoch is kept 2000, for all. Dropout layers are used at appropriate places for avoiding overfitting.

Results are shown in Fig. 2. Neural network with "α" = 2 performs better than "α" = 5 in terms of accuracy. Smaller values of "α" enable more number of neurons per layer, so the results are obvious. CNN performs better than both of them. Pattern of accuracy shows in the first phase of epochs neural networks learn and go on increasing their accuracy. In the second phase, the accuracy is stabilized and will be steady even if the number of epochs increases.

In future, with the notion of IoT based systems, it is expected that a huge data will be generated for the analysis [18]. Under such scenario, techniques that will reduce the dimensionality during analysis will be most preferred for speedy analysis and data processing. PCA and ICA can be the best candidate to process the big data arised from medical applications [19]. Hence, in future, we will extend the present work towards the integration of PCA and ICA based module.

4 Conclusion

In this work, novel approach to apply CNN to detect diabetes is showcased. Also, fuzzification in matrix form is a proposed here. The entire data set is fuzzified and thus it gets multiple values for each of the feature. In other words, each attribute got different membership in different classes, and thus the entire data set is populated. So data is represented in a form of matrix of fuzzy values.

Here, three different experiments have been conducted. The first two experiments were on classical neural networks with value of "α" = 2 and "α" = 5 and then we compare the results with convolutional network with fuzzified input. It is found that using CNN with fuzzification for diabetes detection is better than normal NN.

References

1. American Diabetes Association: Diagnosis and classification of diabetes mellitus. Diabetes Care **37**(1), s81–s90 (2014)
2. Diagnosis and classification of diabetes mellitus: new criteria. Am. Fam. Physician **58**(6), 1355–1362 (1998)
3. Ross, T.: Fuzzy Logic with Engineering Applications, 3 edn. Wiley Student Edition
4. Kumar, S.: Neural Networks: A Classroom Approach, 2 edn. Tata McGraw-Hill Education (2012)
5. Rajasekaran, S., Pai, G.A.V.: Neural Networks, Fuzzy Logic, and Genetic Algorithms Synthesis and Applications, 8 edn. PHI (2003)
6. Gulshan, V., Peng, L., Coram, M.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**(22), 2402–2410 (2016)
7. Pinheiro, P., Collobert, R.: Recurrent convolutional neural networks for scene labeling. Proc. Mach. Learn. Res. **32**(1), 82–90 (2014)
8. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
9. Kamble, T.P., Patil, S.T.: Diabetes detection using deep learning approach. Int. J. Innov. Res. Sci. Technol. **2**(12), 342–349 (2016)
10. Coenen, F., Broadbent, D.M., Harding, S.P., Zheng, Y., Pratt, H.: Convolutional neural networks for diabetic retinopathy. Procedia Comput. Sci. (2016)
11. Yamini, C., Tharani, S.: Classification using convolutional neural network for heart and diabetics datasets. Int. J. Adv. Res. Comput. Commun. Eng. **5**(12), 417–422 (2016)
12. Asati, V.C.S.: Automatic detection of diabetic retinopathy using deep convolutional neural network. Int. J. Adv. Res. Ideas Innov. Technol. **3**, 633–641 (2017)
13. Liu, K., Lai, S., Zhou, G., Zhao, J., Zeng, D.: Relation classification via convolutional deep neural network. In: National Laboratory of Pattern Classification (2014)
14. Karen Simonyan, Z.: Very deep convolutional neural network for large scale image recognition. ICLR (2015)
15. Fadewar, H.S., Deshmukh, T.: Machine predicts the diagnosis a brief review of medical diagnosis by machine learning techniques. Indian J. Comput. Sci. Eng. **8**(5), 636–638 (2017)
16. Fadewar, H., Deshmukh, T.: Data mining techniques for diagnosis of diabetes: a review. Int. J. Emerg. Res. Manag. Technol. **6**(9), 212–214 (2017)
17. Lichman, M. http://archive.ics.uci.edu/ml (2013)
18. Deshpande, P., Iyer, B.: Research directions in the Internet of Every Things (IoET). In: International Conference on Computing, Communication and Automation (ICCCA), pp. 1353–1357 (2017)

19. Patil, M., Iyer, B., Arya, R.: Performance evaluation of PCA and ICA algorithm for facial expression recognition application. In: Proceedings of Fifth International Conference on Soft Computing for Problem Solving, pp. 965–976 (2016)
20. Chandore, V., Asati, S.: Automatic detection of diabetic retinopathy using deep convolutional neural network. Int. J. Adv. Res. Ideas Innov. Technol. **3**(4), 633–641 (2017)
21. https://deeplearning4j.org/neuralnet-overview. (2017)

Classification of Magnetic Resonance Brain Images Using Local Binary Pattern as Input to Minimal Complexity Machine

Heena Hooda and Om Prakash Verma

Abstract Magnetic Resonance Imaging (MRI) is a powerful visualization tool that is extensively used in medical laboratories to capture images of internal anatomy of human body. Classification of MRI brain images into tumorous and non-tumorous image is a critical and time-consuming task for the radiologist. Correct and computerized classification of MRI brain images is very important for their investigation and analysis. In this paper, we have proposed to use binary patterns (LBP) as features to classify MRI brain images into tumorous and non-tumorous. The LBP computes the relationship between central pixel and neighboring pixels of the 3×3 window and assigns a label to each window. The histogram of these labels is then used as a feature vector that is fed into the classification stage. The images are classified using Minimal complexity machine (MCM) algorithm. As compared to Support Vector Machine (SVM) algorithm, MCM performs better generalization and makes use of lesser number of support vectors. The performance analysis of the proposed techniques is done on the basis of accuracy calculated, and it is found that the classification rate is superior to other existing algorithms.

Keywords Brain image segmentation · Local binary pattern
Minimal complexity machine

1 Introduction

Medical image processing is one of the most exigent and emerging fields nowadays. The technique used to capture scans of the human body for examination is known as medical imaging. Magnetic resonance imaging (MRI) is one of the techniques of imaging technology which is widely used for high-quality medical images [1, 2].

H. Hooda (✉) · O. P. Verma
Department of IT, Delhi Technological University, New Delhi 011042, India
e-mail: heenahooda@gmail.com

O. P. Verma
e-mail: opverma.dce@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_90

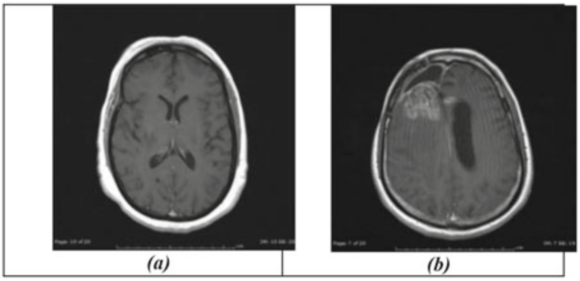


Fig. 1 **a** Non-tumorous, **b** tumorous MRI slices of brain

MRI provides most detailed anatomical information and it does not make use of high energy radiation as compared to CT scan, hence it is a safer method of imaging. MRI is bleeding, swelling, and discovers other such abnormalities of brain. The growth of abnormal cells in the tissues of the brain is known as brain tumor as shown in Fig. 1, and it is the second dominant reason of tumor death. Fully automatic tumorous and non-tumorous human brain image classification can be achieved from magnetic resonance images that are of utmost significance for research and clinical studies. The MR images for tumorous and non-tumorous brain are shown in Fig. 1.

Recent studies have shown that magnetic resonance brain images can be classified using supervised techniques such as vector machine [3], and unsupervised classification techniques such as self-organization map (SOM) [4] and fuzzy *c*-means (FCM) [5]. A survey paper on Computer-aided diagnosis of human brain tumor through MRI is given by Dahshan et al. in [6] and a new algorithm is also discussed that consists of three stages, namely segmentation, features extraction and reduction, and classification of MR images into normal or abnormal. In 2011, a hybrid method for MRI brain image classification is given by Zhang et al. [7] which utilize a neural network (NN)-based method to classify a given MR brain image into normal or abnormal brain. Hybrid intelligent techniques are used for MRI brain images classification. Dahshan et al. [8] present a hybrid technique for the classification of MRI images. Two classifiers have been proposed by the author in the classification stage. The first classifier is based on feed-forward back-propagation artificial neural network (FP-ANN) and the second classifier is based on *k*-nearest neighbor (KNN). Classification of magnetic resonance brain images using wavelets as input to SVM and NN is proposed by Chaplot et al. [4] in which a novel method using wavelets as input to NN is used, and SOM and SVM is implemented for classification of magnetic resonance (MR) images of the human brain. The feature extraction and reduction needs to be carried out for classification of MR images. The features of the brain image can be extracted using a discrete wavelet transform (DWT) as done by Hema Rajini et al. in [9]. Nandpuru et al. [10] have extracted features using gray-level co-occurrence matrix (GLCM) and further classified them using SVM. Feature reduction can be carried out using principal component analysis (PCA) as illustrated by Ibrahim et al. in [11]. In [12], Sridhar et al. proposed an algorithm for extraction of features using Discrete Cosine Transform (DCT) and classifying the images with the help of Probabilistic Neural Network (PNN). A hybrid method is given by Yazdani et al. [13] by integrating the modified Expectation Maximization (EM) and GLCM features and feeding them for classification into SVM algorithm. Also, a hybrid classifier [14] was given by Macchale et al. by integrating the features of KNN and SVM classifier. Gupta et al. proposed an algorithm for multi sequential image classification of brain images by using (Effective Information Feature Extraction) EIFE and SVM [15].

Many techniques for extraction of features from brain images have already been proposed but still there is a scope to investigate some novel feature extraction technique. Sorensen et al. [16] have proposed to use LBP as features in a classification framework for classifying different texture patterns in lung computed tomography. But the application of LBP features to brain images has not yet been explored. In this paper, we first discuss the methodology in Sect. 2. The process for feature extraction using LBP is described in Sect. 3. Then, Sect. 4 explain various techniques for classification. The experimental results and comparison of the classification algorithms used are done in Sect. 5. The paper is concluded in Sect. 6.

2 Methodology

The goal of image classification is to group pixels having similar features into one class depending on the features obtained. The process of feature extraction is carried out using LBP that has been explained in this section. The proposed system for classifying MRI brain images into two different classes is shown in Fig. 2. The techniques used for classification can be either supervised or unsupervised. As we can see from the figure, the classification system is constituted of two stages, namely training stage and testing stage. In this paper, we have applied MCM algorithm for which classification offers various advantages over the well-known SVM and KNN algorithm.

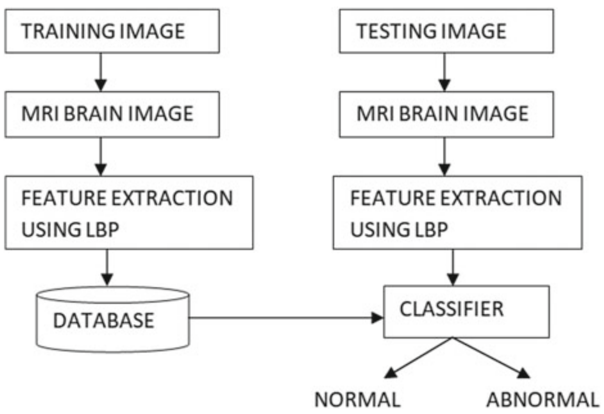


Fig. 2 Brain image classification system

3 Feature Extraction

In this paper, we use LBP for extracting features from MRI brain sample images. Originally, LBP was proposed by Ojala et al. [17] for deriving the textural features from the image. In his basic LBP method, every pixel of the image is converted into a binary number on the basis of thresholding the 3×3 window which comprises of the neighbors of the pixel. The thresholding is done by comparing the value of the center pixel with the neighboring pixels. This thresholding results in a binary number and these binary values are concatenated by moving in a clockwise direction starting from top-left neighbor as shown in Fig. 3. The binary codes are converted into decimal number to label the pixel and are termed as LBPs or LBP codes. The steps for constructing LBP codes are stated in algorithm 1.

The binary number thus formed is said to be uniform, if the pattern consists of two or less than two transitions between consecutive bits from 0 to 1 or 1 to 0 considered circularly. Thus, we can see that the binary string 00000000, 001110000, and 11100001 are uniform patterns having zero, two, and two transitions, respectively, while 11001001 and 01010011 are nonuniform with four and five transitions, respectively. These uniform patterns carry around 90% of information in (8, 1) neighborhood [18]. Ahonen et al. [19] reported that 90.6% of patterns in (8, 1) neighborhood are uniform while experimenting on FERET database. The concept of uniform patterns paved the way to a much-reduced feature vector by storing the uniform and nonuniform pattern in two different bins. Following this criterion, the histogram reduced to 59 bins in case of uniform LBP against 256 bins in standard LBP. The histogram of the LBP labeled image can be illustrated as in Eq. 1 [20].

$$H_i = \sum_{x,y} I(f(x,y) = i), \quad i = 0, 1, \ldots, n-1 \tag{1}$$

This histogram forms the feature vector having local micro-pattern information like edges, spots, and flat areas over the entire image [20]. Thus, a statistical representation of the image is obtained. The final, spatially enhanced, and global feature vector is derived by joining all the region specific histograms and is shown in Eq. 2:

$$H_{i,j} = \sum_{x,y} I(f(x,y) = i) I\big((x,y) \in R_j\big)\, i = 0, 1, \ldots, n-1 and j = 0, 1, \ldots, m-1 \tag{2}$$

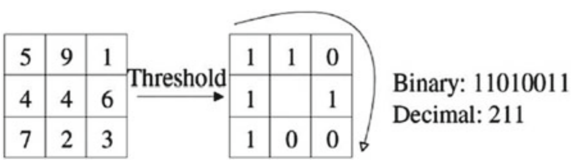


Fig. 3 The basic LBP operator [24]

Algorithm 1 Steps for Constructing LBP Code

1. Take a 3×3 window.
2. In the window, take central pixel and threshold its eight neighbors by comparing.
3. If the intensity of the center pixel is greater than or equal to its neighbor, then denote it with 1 and 0 if not.
4. The result will be a 8-bit binary number for each pixel, just like 11001111.

4 Image Classification

The techniques used for image classification step can be either supervised or unsupervised. In this section, we discuss some supervised image classification algorithms, namely KNN and SVM along with their advantages and disadvantages. A novel MCM algorithm for classification of MRI brain images which also belongs to the category of supervised classification have been discussed briefly in this section.

4.1 k-Nearest Neighbor

It is the simplest classification algorithm, which is used to classify the data in the sample to one of the classes in the training by implementing the nearest neighbor technique. A distance measure is used as a metric to classify the input feature vector V into classes on the basis of training vectors which are closest to the data. The vector I belongs to the class to which the majority of those *k*-nearest neighbors belong. The KNN algorithm is dependent on two functions calculated using the *k*-nearest neighbors, namely distance function and voting function [7]. The metric for distance function can be Euclidean, cosine similarity, correlation or hamming distance, and Mahalanobis distance. The KNN classification algorithm is a traditional technique, which belongs to a class of supervised algorithms and is not based on any parameters except the value of K. The algorithm works efficiently for optimal values of K.

Similar to other learning algorithms, KNN also comprises of two stages, namely training stage and testing stage. First, in the training stage, data points are allocated in an n-dimensional space. Each training data point is assigned with a label to indicate the class it fits in. Then, in the testing stage, the unlabeled data points are tested for the identification of class. The class is identified by producing a list of K-nearest data points to the input data point by making use of the already classified data. The algorithm then returns the class to which the majority of the labeled data belongs. The value of K given as input to the algorithm plays an important role in accurately classifying the data.

4.2 Support Vector Machine

SVM is a binary classification algorithm that takes labeled data as input along with the group or class it belong to and outputs the classes for the unlabeled data on the basis of training. The SVM algorithm given by Vapnik [21] was developed from the idea of minimizing the structural risk. SVM are generally used to classify the data into two classes, and are found to be more attractive and systematic as compared to KNN in learning to build a linear or nonlinear boundary between classes. The SVM algorithm also involves two phases, training and testing like any other machine learning algorithm. In the training phase, the known data is fed as input to the SVM along with the correct labels that are known prior. The training phase makes the SVM intelligent to accurately classify the unknown data. SVM classifier makes a hyperplane between the two classes which can be either linear or nonlinear [4]. Let there be a data set $x \in R^n$ which needs to classified and it belongs to class label denoted by $y = \pm 1$.

In linear SVM classifier, the input data or images can be separated linearly. It is the most simple case of data classification and linear function used to perform it is given by the equation of the form:

$$f(x) = W^T x + b \tag{3}$$

such that for each training data or image x_i, the function yields $f(x_i) \geq 0$ for y_i = +1 and $f(x_i) < 0$ for $y_i = -1$. Hence, training data from the two different classes are separated by the hyperplane given by the equation

$$f(x) = W^T x + b = 0 \tag{4}$$

For a given set of data set, many hyperplanes exist to distinguish between the two classes. The SVM classification algorithm chooses the hyperplane in which the distance of the separating margin between the two groups or classes is maximum.

The nonlinear classifier is an extension of linear classifier where the input data or image x, is mapped into a higher dimensional space using a nonlinear operator $\varphi(.)$. The nonlinear classifier obtained is given by the equation

$$f(x) = W^T \varphi(x) + b, \tag{5}$$

which is linear in terms of the transformed data $\varphi(\mathrm{x})$ but nonlinear in terms of the original data $x_i \in R^n$.

The kernel function used in the SVM algorithm plays a major role of inevitably transforming the input vector into a high dimensional feature space. As we know that all real-world problems cannot be linearly separated into two classes, hence, a linear SVM classifier cannot be used for all the cases. Therefore, a need arises for introduction of kernel function which permits the data points to enter a class belonging to different class label. However, when choosing a kernel function, it is necessary to check whether it is associated with the inner product of some nonlinear mapping. Various kernel functions that are used in the SVM algorithm are polynomial, Radial Basis Function (RBF), and Multilayer Perceptron (MLP) kernel. During implementation, we need to initialize the parameters of each kernel. The order of the polynomial of kernel function should be given initially during the training phase. The value of sigma needs to be specified for the RBF. The MLP kernel requires two parameters, P1 and P2, where P1 > 0 and P2 < 0.

4.3 Minimal Complexity Machine

The learning capacity of the machine is calculated with the help of Vapnik–Chervonenkis (VC) dimension and computational theory shows that good generalization and robust learning can be achieved using a small VC dimension. Since SVM can have large and infinite VC dimension they do not provide the guarantee of good generalization. The solution to this is proposed by Jayadeva in [22] by finding a hyperplane classifier with small or minimum VC dimension. This algorithm is termed to be the MCM algorithm. It is shown to be performing better generalization as compared to SVM and uses less than one-tenth the number of support vectors. In this paper, we have used the MCM algorithm to classify brain MRI images. Like any other machine learning algorithm, MCM also involves two basic steps, known as training step and testing step. In the training step, the known data is fed into the MCM classifier along with the previously known outputs. The classifier gains its intelligence from the training step.

The soft margin equivalent of the linear MCM is given by

$$\min_{w,b,h} \mathrm{h} \ + \ \mathrm{C} \ . \sum_{i=1}^{M} q_i \tag{6}$$

where,

$$h = \frac{\max_{i=1,2,\ldots,M} \left\| u^T x^i + \mathrm{v} \right\|}{\min_{i=1,2,\ldots,M} \left\| u^T x^i + \mathrm{v} \right\|} \tag{7}$$

$$h \ \geq \mathrm{y}_i.\ [\mathrm{w}^T x^i + \mathrm{b}] + \mathrm{q}_i, \quad \mathrm{i} = 1, 2, \ \ldots, \ \mathrm{M} \tag{8}$$

$$y_i.\ [\mathrm{w}^T \mathrm{x}^i + \mathrm{b}] + \mathrm{q}_i \ \geq 1, \quad \mathrm{i} \ = 1, 2, \ldots, \ \mathrm{M} \tag{9}$$

$$q_i \ \geq 0, \quad \mathrm{i} \ = 1, 2, \ldots, \ \mathrm{M}. \tag{10}$$

Once w and b have been determined by solving (6)–(10), the class of a test sample x may be determined from the sign of the discriminant function

$$f(x) = W^T x + b = 0 \tag{11}$$

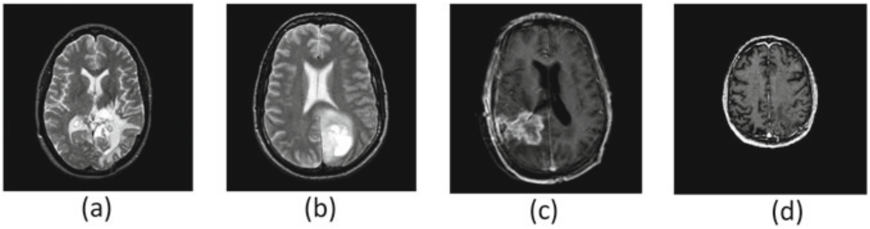


Fig. 4 Sample of brain MRI Images: **a**–**c** abnormal brain; **d** normal brain (http://www.med.harvard.edu/AANLIB/)

Here, the value of C allows a trade-off between the machine capacity of the classifier and the classification error. Once w and b have been determined, the class of a test sample x may be determined as before by using the sign of f(x) in Eq. 10.

The corresponding optimization problem for the kernel MCM may be shown to be

$$\min_{w,b,h,q} \mathrm{h} + \mathrm{C} \,.\, \sum_{i=1}^{M} q_i \tag{12}$$

$$h \geq \mathrm{y}_i.\, [\mathrm{w}^T \phi(x^i) + \mathrm{b}] + \mathrm{q}_i, \quad \mathrm{i} = 1, 2, \ldots, \mathrm{M} \tag{13}$$

$$y_i.\, [\mathrm{w}^T \phi(\mathrm{x}^i) + \mathrm{b}] + \mathrm{q}_i \geq 1, \quad \mathrm{i} = 1, 2, \ldots, \mathrm{M} \tag{14}$$

$$q_i \geq 0, \quad \mathrm{i} = 1, 2, \ldots, \mathrm{M}. \tag{15}$$

5 Experimental Results

The algorithm is implemented on personal computer (Intel Core i7 @ 2.0 GHz CPU, 8 GB RAM) using MATLAB 7.9.0 (2009b). The datasets consist of T2-weighted MRI brain images in axial plane and 256 × 256 in plane resolution. The images are downloaded from the Harvard Medical School Website (http://www.med.harvard.edu/AANLIB/) [23]. Some of the sample images are shown in Fig. 4. The dataset consists of 80 T2-weighted axial MRI brain images of which 24 are of normal and 56 are of abnormal brain. The experiments are also conducted on real-time database taken from Rajiv Gandhi Cancer Institute and Research Centre (RGCI&RC), Delhi as shown in Fig. 5. To check the accuracy of the algorithms, the system is first fed with the same training and testing set, and the match rate of 100 percent is achieved.

The features from the images are extracted using LBP as explained in Sect. 3 and are given as input to the classification algorithms. The performance of the system is evaluated in terms of accuracy as it measures the overall performance of the system. It is calculated by identifying the correctly classified cases out of all the samples as given in Eq. 16.

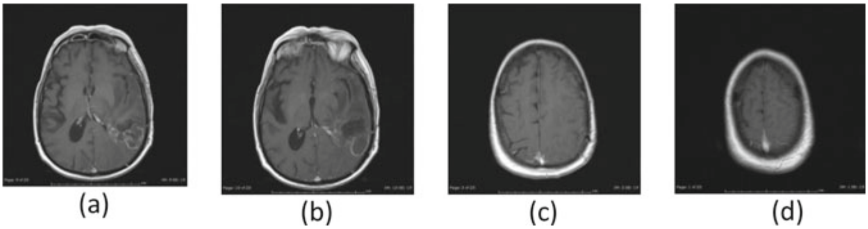


Fig. 5 Sample of brain MRI images from RGCI&RC: **a**, **b** abnormal brain; **c**, **d** normal brain (RGCI&RC)

Table 1 Classification accuracy comparison for the same MRI dataset

Year	Author	Technique	Accuracy (%)
2006	Chaplot et al.	DWT + SOM	86.25
2006	Chaplot et al.	DWT + SVM (linear)	95.00
2006	Chaplot et al.	DWT + SVM (polynomial)	96.25
2006	Chaplot et al.	DWT + SVM (radial basis function)	96.25
2010	Dahshan et al.	DWT + PCA + KNN	93.75
2010	Dahshan et al.	DWT + PCA + ANN	91.25
2011	Zhang et al.	DWT + PCA + BPNN	92.50
2013	Sridhar et al.	DCT + PNN	88.75
2015	Veeramuthu et al.	DWT + GLCM + PNN-RBF	90.00
2015	Yazdani et al.	EM + GLCM + SVM	91.25
2015	Machhale et al.	SVM + KNN	88.75
2017	Gupta et al.	EIFE + PCA + SVM	93.75
2017	Proposed	ULBP + MCM	98.75

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

where TP, FP, TN, and FN denote the number of cases of true positive, false positive, true negative, and false negative, respectively.

The accuracy of the proposed algorithm (LBP + MCM) is compared with the KNN and SVM algorithms using LBP features. We found that LBP features give 90% accuracy when used with KNN, 95% accuracy with SVM (linear), 96.25% with SVM (polynomial and radial basis function). Also, the comparison is done with other brain image classification algorithms existing in the literature as shown in Table 1. The results in Table 1 show that our method outperforms other classification methods in the literature with an accuracy of 98.75%. We can also see the comparative analysis of MRI brain image segmentation techniques in Fig. 6.

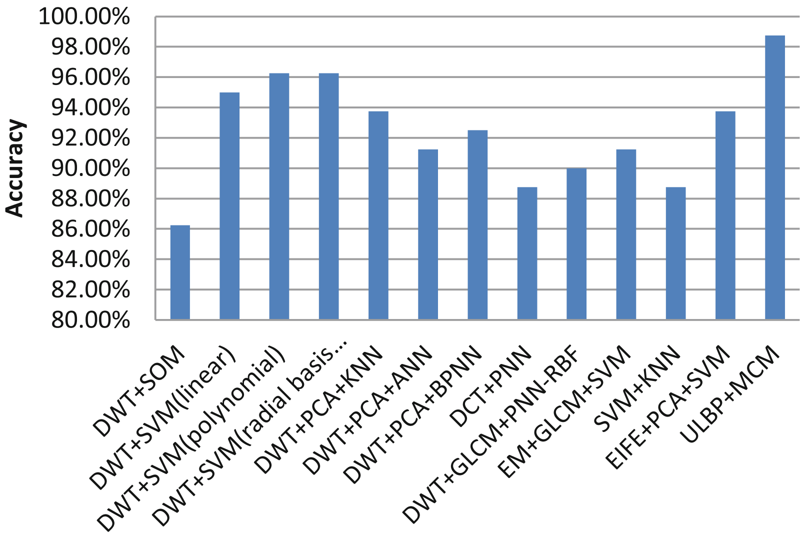


Fig. 6 Comparative analysis of MRI brain image segmentation techniques

6 Conclusion

In this paper, we have proposed a novel feature extraction and classification algorithm for MRI brain images. We have proposed to use LBP as features to classify MRI brain images into tumorous and non-tumorous. The LBP computes the relationship between central pixel and neighboring pixels of the 3×3 window and assigns a label to each window. The histogram of these labels is then used as a feature vector that is fed into the classification stage. The images are classified using MCM algorithm. The performance analysis of the proposed techniques is done on the basis of accuracy calculated and it is observed that the classification rate is higher as compared to other existing algorithms. The method obtained 98.75% classification accuracy on test images of the selected datasets. In future, multiple class classification of brain MRI images should be explored.

References

1. Iftekharuddin, K.M.: On techniques in fractal analysis and their applications in brain. In: Medical imaging systems: technology and applications, Analysis and Computational Methods, vol. 1. World Scientific Publications (2005). ISBN 981-256-993-6
2. Fan, Y., Rao, H., Giannetta, J., Hurt, H., Wang, J., Davatzikos, C., Shen, D.: Diagnosis of Brain Abnormality Using both Structural and Functional MR Images. In: Proceedings of the IEEE Conference on Engineering in Medicine and Biology Society (2006)
3. Abdullah, N., Ngah, U.K., Aziz, S.A.: Image classification of brain MRI using support vector machine. In: IEEE International Conference on Imaging Systems and Techniques (IST), 17–18

May 2011, pp. 242–247
4. Chaplot, S., Patnaik, L.M., Jagannathan, N.R.: Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network. Biomed. Signal Process. Control **1**(1), 86–92 (2006)
5. Nandha Gopal, N., Karnan, M.: Diagnose brain tumor through mri using image processing clustering algorithms such as fuzzy c means along with intelligent optimization techniques. In: IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Dec 2010, pp. 1–4
6. El-Dahshan, E.-S.A., Mohsen, H.M., Revett, K., Salem, A.-B.M.: Computer-aided diagnosis of human brain tumor through MRI: A survey and a new algorithm. Expert Syst. Appl. **41**(11), 5526–5545 (2014)
7. Zhang, Y., Dong, Z., Wua, L., Wanga, S.: A hybrid method for MRI brain image classification. Expert Syst. Appl. **38**(8), 10049–10053 (2011)
8. El-Dahshan, E.-S.A., Hosny, T., Salem, A.-B.M.: Hybrid intelligent techniques for MRI brain images classification. Digit. Signal Process. **20**(2), 433–444 (2010)
9. Rajini, N.H., Bhavani, R.: Classification of MRI brain images using k-nearest neighbor and artificial neural network. IEEE International Conference on Recent Trends in Information Technology (ICRTIT), 3–5 June 2011, pp. 563–568
10. Nandpuru, H.B., Salankar, S.S., Bora, V.R.: MRI brain cancer classification using support vector machine. In: IEEE Students Conference on Electrical, Electronics and Computer Science (SCEECS), 1–2 Mar 2014, pp. 1–6
11. Ibrahim, W.H., Osman, A.A.A., Mohamed, Y.I.: MRI brain image classification using neural networks. In: International Conference on Computing, Electrical and Electronics Engineering (ICCEEE), 26–28 Aug 2013, pp. 253–258
12. Sridhar, D., Murali Krishna, I.: Brain tumor classification using discrete cosine transform and probabilistic neural network. In: 2013 International Conference on Signal Processing, Image Processing & Pattern Recognition, Coimbatore, pp. 92–96 (2013)
13. Yazdani, S., Yusof, R., Pashna, M., Karimian, A.: A hybrid method for brain MRI classification. In: 2015 10th Asian Control Conference (ASCC), Kota Kinabalu, pp. 1–5 (2015)
14. Machhale, K., Nandpuru, H.B., Kapur, V., Kosta, L.: MRI brain cancer classification using hybrid classifier (SVM-KNN). In: 2015 International Conference on Industrial Instrumentation and Control (ICIC), Pune, pp. 60–65 (2015)
15. Gupta, T., Gandhi, T.K., Panigrahi, B.K.: Multi-sequential MR brain image classification for tumor detection, Recent advancements in computer, communication and computational sciences. J. Intell. Fuzzy Syst. **32**(5), 3575–3583 (2017)
16. Sørensen, L., Shaker S.B., de Bruijne, M.: Texture Classification in Lung CT Using Local Binary Patterns. In: Metaxas D., Axel L., Fichtinger G., Székely G. (eds.) Medical Image Computing and Computer-Assisted Intervention—MICCAI 2008. MICCAI 2008. Lecture Notes in Computer Science, vol. 5241. Springer, Berlin, Heidelberg (2008)
17. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recogn. **29**, 51–59 (1996)
18. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**, 971–987 (2002)
19. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **28**, 2037–2041 (2006)
20. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image Vis. Comput. **27**, 803–816 (2009)
21. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
22. Jayadeva: Learning a hyperplane classifier by minimizing an exact bound on the VC dimension. In: Neurocomputing, Part B, 3 Feb 2015, vol. 149, pp. 683–689
23. Harvard Medical School. http://www.med.harvard.edu/AANLIB/
24. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. In: Computer vision-eccv 2004, Springer, pp. 469–481 (2004)

Underwater Image Colour Balance by Grey World Approach with Attenuation Map

Sonali Sankpal and Shraddha Deshpande

Abstract Underwater images are degraded by attenuation of light in water. This attenuation depends upon wavelength and depth in water. One of the effects of degradation of image in water is absorption of colour giving greenish-blue hue to the image. Because of this colour fading of underwater images, colour correction is the first preprocessing step in underwater image processing. Many researchers attempted colour correction methods but most of it operates globally. Global colour correction methods give reddish effect to image. The method proposed in this paper used Grey World approach for colour correction, but it is modified using attenuation map. Use of attenuation map avoids saturation of colours and colour corrects only those pixels which are significantly attenuated. Results of the proposed method are compared with state-of-the-art methods by quality metrics mean square error, structural similarity index and entropy of image. It is seen that the proposed method in this paper gives better results than state-of-the-art methods.

Keywords Grey world · Attenuation map · Colour correction · Underwater image

1 Introduction

Images captured in water are degraded because of interaction of light with water [1]. It is given as

$$E(d) = E(0)e^{-\alpha d} \tag{1}$$

where E(d) is irradiance at distance d, α is attenuation coefficient. The interaction is of two types , absorption and scattering of light in water, so Eq. (1) can be written as

S. Sankpal (✉)
Department of E & TC Engineering, PVPIT, Budhagaon, Sangli, Maharashtra, India
e-mail: sssankpal.etc@pvpitsangli.edu.in

S. Deshpande
Department of Electronics Engineering, WCE, Sangli, Maharashtra, India

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_91

$$E(d) = E(0)e^{-\beta d}e^{-\gamma d} \tag{2}$$

where β and γ are absorption and scattering coefficients, respectively. According to Funk et al. [2], water window is spectral bandwidth for which light penetrates seawater with lowest attenuation. For clear water, this window is centred at 4800A^0. Unlike in air, attenuation in water is wavelength dependent [2]. When light travels through water its colour drops off [1]. First, red colour drops of at the depth of 3 m, orange at 5 m then yellow at 10 m, finally green at 20 m and blue travels longest. As a result of attenuation of colours with depth in water, deepwater images are dominated by greenish-blue hue. Many imaging devices with filters can also attenuate colours. Many underwater applications such as marine archaeology, study of marine eco system, classification of marine habitat, navigation require colour correction of underwater videos or images. Use of artificial light can solve the problem of colour attenuation but introduces different problem of non-uniform illumination. So, underwater imaging applications should work on colour correction process in addition to image enhancement and restoration process.

Rest of the paper is divided into four sections, Sect. 2 described state-of-the-art methods of the problem, Sect. 3 presented the proposed method, Sect. 4 discussed results and Sect. 5 have given concluding remark.

2 Related Work

The very first method found for colour correction of underwater image is proposed by Torres-Mendez and Dudek [3]. They modelled image by Markov Random Field (MRF) where each pixel in input image is assigned a colour value that can be best described by its surroundings using training image patches. Here, training set is composed of ground truth image patches. Vasilescu in [4] use Colour Rendering Index (CRI) to evaluate colour rendering ability of light source, which is computed with some experimental setup and some prior knowledge of a system. Bazeille et al. [5] proposed a simple method where no prior knowledge of any parameter is required, they simply equalised each colour mean. The same method is followed by Çelebi and Ertürk [6]. Chiang and Ying [7] estimated depth of scene and then, the effect of depth are removed from the image. This removal of effect of depth compensates for absorption of colour in image and colour corrects the image. The hypothesis of colour correction by Naim and Isa [8] is, pixel distribution of an image is around diagonal axis of RGB cube. Each colour channel of image is shifted based on white reference point, so white reference point is a crucial point in this colour correction method. Iqbal et al. [9] perform unsupervised colour correction using Von Kries hypothesis. In this hypothesis, the maximum value of red, green and blue channel are calculated, and average value for this channel is also computed as R_{avg}, G_{avg} and B_{avg} and set and set blue average as target value to find multipliers for other two channels. The multiplier for other two channels are calculated to balance the colour in the images as

$$A = \frac{B_{avg}}{R_{avg}} \quad \& \quad B = \frac{B_{avg}}{G_{avg}} \tag{3}$$

The values of inferior colour channels are adjusted by von Kries hypothesis as

$$R' = A \times R \quad \& \quad G' = B \times G \tag{4}$$

where R and G are original pixel values and R′ and G′ are colour-corrected values. Abdul Ghani and Isa [10] modified von Kries hypothesis to colour correct the underwater image. In modified von Kries hypothesis, colour average of median colour channel is set as target mean and multipliers for other two colour channels are calculated accordingly as given below

$$A = \frac{median\left(R_{avg}, G_{avg}, B_{avg}\right)}{\min\left(R_{avg}, G_{avg}, B_{avg}\right)} \tag{5}$$

$$B = \frac{median\left(R_{avg}, G_{avg}, B_{avg}\right)}{\max\left(R_{avg}, G_{avg}, B_{avg}\right)} \tag{6}$$

In this method, the colour channel with minimum intensity value is multiplied with A from Eq. (5) and with maximum intensity value is multiplied with B from Eq. (6). Bianco [11] presented colour correction of underwater images using $l\alpha\beta$ colour space. It uses grey world hypothesis and subtracts mean of α and β from α and β channels, respectively, and then image is converted from $l\alpha\beta$ colour space to RGB colour space. Abdul Ghani and Isa [12] discussed histogram matching method where they matched histogram of red channel to histogram of blue channel for colour correction. Ancuti et al. [13] proposed a method in which they colour transfer the image with global statistics and blended with an original image with single-scale fusion. Here, images are guided by weight map which is calculated by averaging red channel with saliency weight, where weight map values are decreases with depth. Fu [14] and Li et al. [15] found maximum and minimum deviations for each colour channel, and then stretch intensity values accordingly. Ancuti et al. [16] white balanced the image by Grey-World algorithm. This method improves red colour only in those regions where the red colour is attenuated the most. They also assume that the green channel is well preserved in underwater images.

With the review of related literature, it is seen that almost all the colour correction techniques work globally except some like the method by Ancuti et al. [13] and Ancuti et al. [16]. The method by C. O. Ancuti et al. [13] consider that red channel is correlated with the attenuation in water, whereas Ancuti et al. [16] consider that green channel is well preserved in water, i.e. both the methods relay totally on a single channel. But it is very difficult to estimate attenuation with single colour channel as there may be some objects with same colours both in foreground as well as background.

3 Proposed Method

According to Ancuti et al. [16] performs better for distorted underwater scene. Traditional grey World approach [17] changes colour globally but there is need of modification for underwater images such that it improves colour in the region where it is necessary. The proposed method in the paper makes a combination of Grey World approach and attenuation map for colour balancing.

3.1 Grey World Approach

The Grey World approach assumes that on average the real world tends to grey. Using this approach, the colour fading in underwater image can be compensated by calculating gain for three colour channels to make the image achromatic. Initially, the mean values for three colour channels are calculated as

$$\begin{aligned} R_{mean} &= mean(R) \\ G_{mean} &= mean(G) \\ B_{mean} &= mean(B) \end{aligned} \tag{7}$$

Here, R_{mean}, G_{mean} and B_{mean} values are mean values for red, green and blue channels, respectively. R, G and B are original colour channels. After calculating mean value using Eq. (7), the target mean value is calculated as

$$\mathrm{max} - \mathrm{mean} = \max(R_{mean}, G_{mean}, B_{mean}) \tag{8}$$

Gain values for three colour channels are calculated using target mean given in Eq. (8) as

$$\begin{aligned} G_R &= \frac{R_{mean}}{max - mean} \\ G_G &= \frac{G_{mean}}{max - mean} \\ G_B &= \frac{B_{mean}}{max - mean} \end{aligned} \tag{9}$$

Here, G_R, G_G and G_B are gain values for red, green and blue channels, respectively.

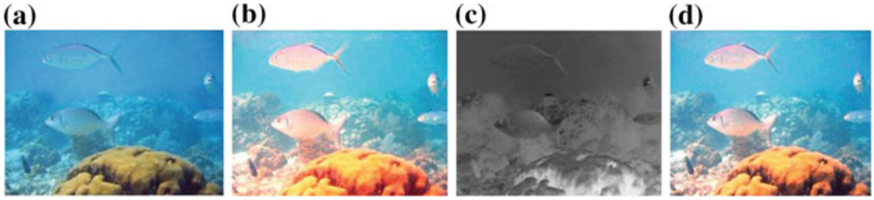


Fig. 1 Colour-corrected image **a** Degraded Image, **b** Colour-corrected by traditional Grey World approach, **c** Attenuation map **d** Colour balanced image by the proposed method

Three colour channels are modified using three gain values to equalise mean values

$$\begin{aligned} R' &= R \times G_R \\ G' &= G \times G_G \\ B' &= B \times G_B \end{aligned} \tag{10}$$

where R′, G′ and B′ are modified values of red, green and blue colour channels, respectively. Underwater images are dominated by blue colour with highest attenuation for red colour, so the value of G_R is very high compared with other gain values. The colour attenuation in water not only varies with wavelength but also varies with depth. But if pixel values for three colour channels are modified using Eq. (10), all pixel values in the image are modified equally, independent of attenuation. This modifies the image to make it reddish in colour. There may be some objects in foreground region, which are already with significant red colour and need not be modified.

So, it is necessary to avoid saturation of red colour as shown in Fig. 1b. and modification of red colour (or any other colour) has to be performed wherever necessary.

3.2 Attenuation Map

The Grey World approach is slightly modified for underwater colour correction. In the new approach, the attenuation of colour in water is considered to modify colour channels. A method by He et al. [18] is used to estimate attenuation map a(x) as

$$a(x) = 1 - k \min_c \left(\min_{y \in \rho(x)} \left(\frac{I_c(y)}{A_c} \right) \right) \tag{11}$$

where x is pixel in image, k is constant and equal to 0.95, c is colour channel and belong to red, green or blue, i.e. $c \in \{R, G, B\}$, ρ is patch and 15×15 patch size gives good results [18]. $I_c(y)$ is intensity value in each colour channel in the patch, whereas A_c is airlight and it is maximum intensity value in the image among brighter pixels in the dark channel. For underwater images, some modification is needed in attenuation map too. When Eq. (11) is applied to underwater images, while performing minimum operation, its value will be zero for most of the images because of high attenuation of red colour. Carlevaris-Bianco et al. [19] suggested a solution to this problem as we consider c ϵ {G, B}, i.e. minimum operation should be performed on green and blue channel without considering red channel. Attenuation map for the image given in Fig. 1a is shown in Fig. 1c.

3.3 Algorithm for Proposed Method

The method proposed in this paper uses the Grey World approach to colour balance the underwater image but slightly modify traditional approach using attenuation map. The algorithm for the method is given as follows:

i. Take I the input image which is colour faded because of underwater environment.
ii. Separate three colour channels R, G and B and find mean values R_{mean}, G_{mean} and B_{mean} as given by Eq. (7).
iii. Find maximum mean value among the three colour channels and set it as target mean.
iv. Find gain value for each colour channel as G_R, G_G and G_B using Eq. (9).
v. Modify three colour channels using the calculated gain as given in Eq. (10). It gives modified image I′.
vi. Find attenuation map a(x) as given by Eq. (11) but with modification suggested by Carlevaris-Bianco et al. [19].
vii. Finally, the colour balanced image I_{cc} is given by Eq (12).

$$I_{cc}(x) = a(x) \times I'(x) + (1 - a(x)) \times I(x) \tag{12}$$

The result of this method is shown in Fig. 1. Input image I is given in (a). Image modified by Grey Word approach I′ is given in (b). Attenuation map a(x) is given in (c). Final colour balanced image I_{cc} after combination of Grey Word approach and attenuation map is given in (d).

4 Results and Discussion

The results of the proposed method are compared with the state-of-the-art methods as discussed in the Sect. 2. The comparison is of two types qualitative and quantitative.

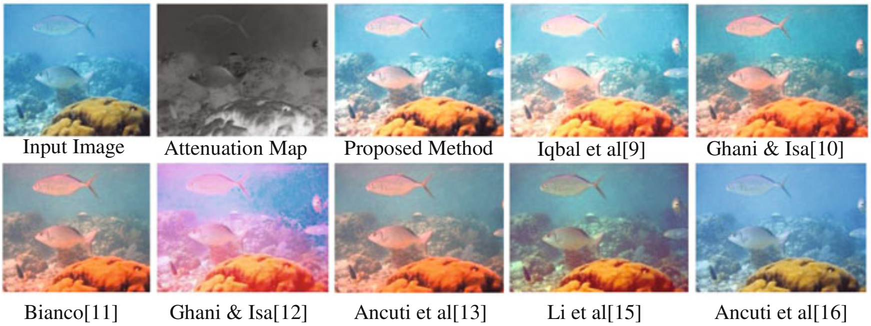


Fig. 2 Results for image 1

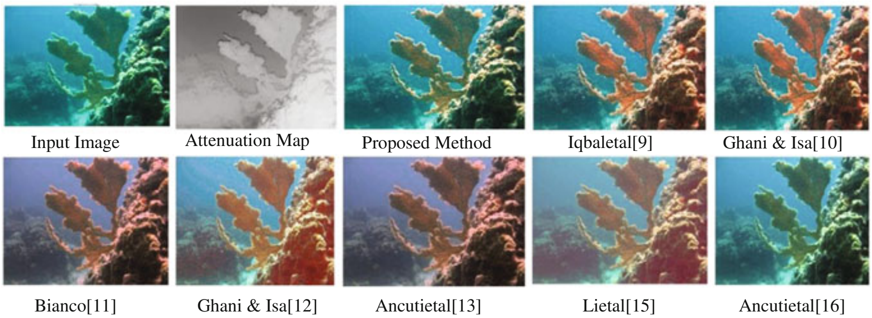


Fig. 3 Results for image 2

4.1 *Qualitative Results*

The results for four images are shown in Figs. 2, 3, 4 and 5. In the results by Iqbal et al. [9] and Ghani and Isa [10], saturation of red colour is seen in some areas of image, which is not observed in the results by the proposed method. Methods by Bianco [11], Ghani and Isa [12] and Ancuti et al. [13] applied a global colour correction, which results in reddish colour cast of image. In the method by Ancuti et al. [16] have not seen a significant improvement in colour. Li et al. [15] and the proposed method are applied considering some local features. But it is difficult to say which method is best because the qualitative comparison is subjective.

4.2 *Quantitative Results*

In case of underwater image processing, ground truth was scarcely available, so it was difficult to compare the results of different algorithms. The solution was devised by Duarte et al. [20] by proposing a dataset with a ground truth. Using this dataset

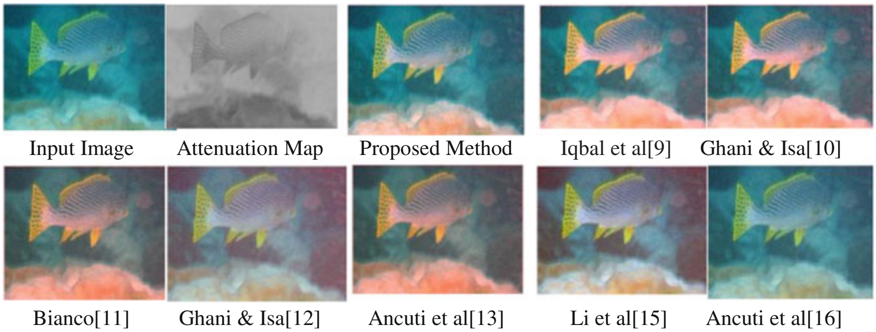


Fig. 4 Results for image 3

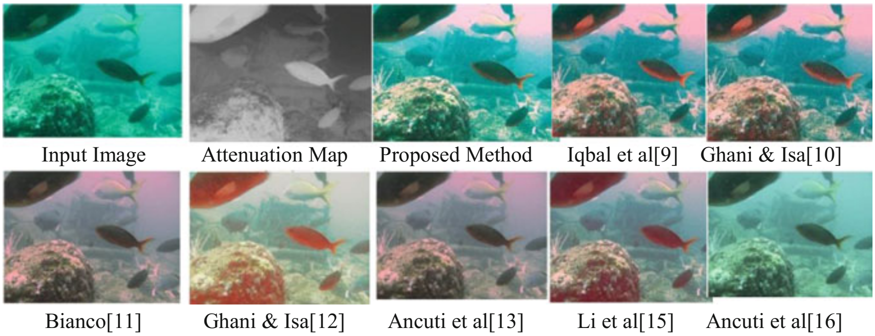


Fig. 5 Results for image 4

now it is possible to compare the results of different methods. In this paper, Deep Blue dataset by Duarte et al. [20] is used for comparison of the results obtained by different methods. The metrics used for comparison are mean square error, structural similarity index [21] and entropy. Mean square error is calculated between ground truth image $I_1(x, y)$ and processed image $I_2(x, y)$ as

$$E = \frac{1}{M \times N} \sum_{x=1}^{M} \sum_{y=1}^{N} [I_1(x, y) - I_2(x, y)]^2 \tag{13}$$

where M and N are rows and columns, respectively. Mean square error is calculated for each colour channel by using Eq. (13) and its' average MSE is given as

$$MSE = \frac{1}{3} \sum_{c \in \{R,G,B\}} E_c \tag{14}$$

Entropy is calculated as

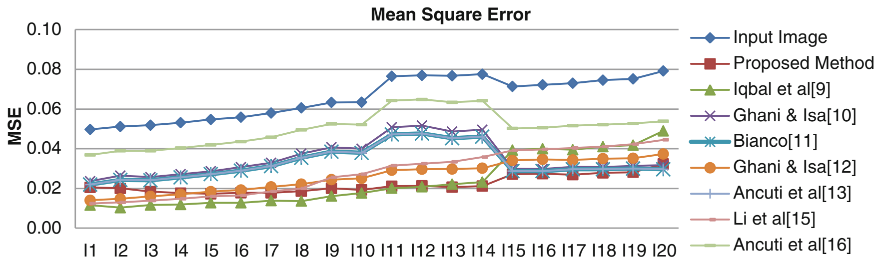


Fig. 6 Graph for mean square error

Table 1 Average values for quality metrics

Methods	MSE	SSIM	Entropy
Input image	0.0658	0.4217	4.5535
Proposed method	**0.0221**	0.5192	**5.5549**
Iqbal et al. [9]	0.0235	**0.5217**	4.5471
Ghani and Isa [10]	0.0349	0.4924	4.2727
Bianco [11]	0.0329	0.5145	4.7473
Ghani and Isa [12]	0.0261	0.5161	4.5525
Ancuti et al. [13]	0.0329	0.5145	4.7473
Li et al. [15]	0.0279	0.5151	4.5232
Ancuti et al. [16]	0.0504	0.4565	4.6408

$$Entropy = \sum p(x) \log_2 p(x) \tag{15}$$

where p(x) is pixel value.

The third quality parameter is SSIM [21] based on structural information of the scene. The results of state-of-the-art methods are compared with results of the proposed method in paper using above three quality metrics for 20 images (I1–I20) of Deep Blue dataset by Duarte et al. [20].

In the dataset from image I1 to image I20, degradation of image increases. Among the quality metrics, smaller the MSE value better the image quality, whereas larger the SSIM and Entropy values, better the quality of image. It is seen from Fig. 6 that mean square error improves as degradation goes on increasing when compared to other methods. For structural similarity index shown in Fig. 7, mixed performance is seen for the proposed method and for some images, it is better than one method whereas for other images, it is better than some other methods. For entropy shown in Fig. 8, the performance of proposed method is best. To evaluate overall performance for all 20 images in dataset, the average values for all the metrics is calculated for all 20 images which are given in Table 1.

In the table, superior value is written in bold type cast.

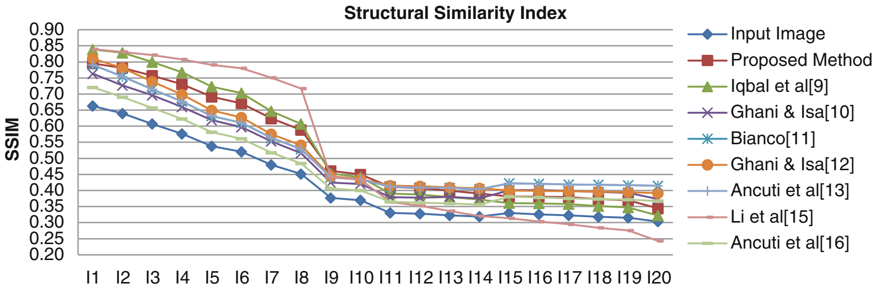


Fig. 7 Graph for structural similarity index

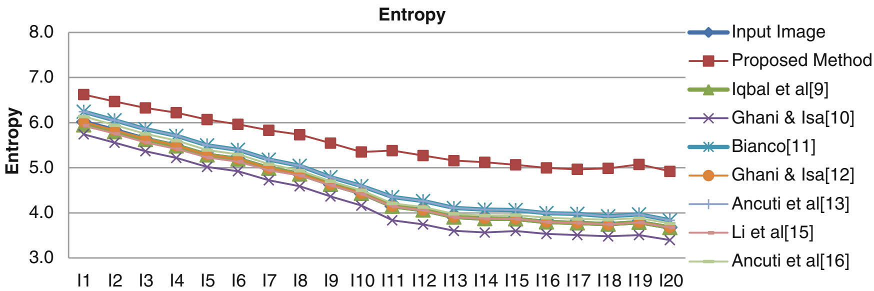


Fig. 8 Graph for entropy

5 Conclusion

In underwater imaging, attenuation in water degrades the image. The degradation not only depends upon the wavelength, but also depends upon the distance the light travels through the water. One of the major degradations is colour fading. The solution to the colour fading by Grey World approach performs global correction, which makes the image reddish. So with this approach, dominant blue colour image is converted into dominant red colour. Many researchers apply the global correction. This paper proposed the method which uses attenuation map with Grey World approach. After qualitative comparison it is observed that global reddish hue present in results of other methods was not seen in the proposed method. In the quantitative comparison, MSE is smallest for proposed method, i.e. method performed best. Entropy value is much better than other methods as the average entropy with the proposed method is closest to original entropy (6.9719) than any other method. For structural similarity, the index value is slightly smaller than the best value, but very close to it. So, we can say that the overall performance considering qualitative, as well as quantitative results for proposed method, is better than other state-of-the-art methods.

References

1. Schettini, R., Corchs, S.: Underwater image processing: state of the art of restoration and image enhancement methods. EURASIP J. Adv. in Signal Process. **2010**, Article ID 746052 (2010)
2. Funk, C., Bryant, S., Heckman, P.: Handbook of underwater imaging system design. Technical Report TP303, Naval Undersea Center, San Diego, Calif, USA (1972)
3. Torres-Mendez, L.A., Dudek, G.: Color correction of underwater images for aquatic robot inspection. In: Proceedings of the 5th International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR '05) (2005)
4. Vasilescu, I., Detweiler, C., Rus, D.: Color-accurate underwater imaging using perceptual adaptive illumination. Auton Robot **31**, 285–296 (2011)
5. Bazeille, S., Quidu, I., Jaulin, L., Malkasse, J.P.: Automatic underwater image pre-processing. In: Proceedings of the Caracterisation du Milieu Marin (CMM '06) (2006)
6. Çelebi, A.T., Ertürk, S.: Visual enhancement of underwater images using empirical mode decomposition. Expert Syst. Appl. **39**(1), 800–805 (2012)
7. Chiang, J.Y., Ying: Underwater image enhancement by wavelength compensation and image dehazing. IEEE Trans. Image Process. **21**(4), 1756–1769 (2012)
8. Naim, M.J.N.M., Isa, N.A.M.: Pixel distribution shifting color correction for digital color images. J. Appl. Soft Comput. **12**(9), 2948–2962 (2012)
9. Iqbal, K., Michael Odetayo, R., James, A., Salam, A., Talib, A.Z.: Enhancing the low quality images using unsupervised colour correction method. In: Proceedings of the IEEE conference on Systems Man and Cybernetics (SMC), pp. 1703–1709, Istanbul, Turkey, Oct 2010
10. Abdul Ghani, A.S., Isa, N.A.M.: Underwater image quality enhancement through composition of dual-intensity images and Rayleigh-stretching, vol. 3, article 757. SpringerPlus (2014)
11. Bianco, G., Muzzupappa, M., Brunoa, F., Garciab, R., Neumann, L.: A new color correction method for underwater imaging, the international archives of the photogrammetry. Remote Sens. Spat. Inf. Sci. **XL-5/W5** (2015); Underwater 3D Recording and Modelling, Piano di Sorrento, Italy, pp. 16–17, Apr 2015
12. Abdul Ghani, A.S., Isa, N.A.M.: Homomorphic filtering with image fusion for enhancement of detail and homogeneous contrast of underwater image. Indian J. Geo-Mar. Sci. **44**(2), 1904–1919 (2015)
13. Ancuti, C.O., Ancuti, C., De Vleeschouwer, C., Garcia, R.: A semi-global colour correction for underwater image restoration. In: Proceedings of SIGGRAPH 17, Posters, Los Angeles, CA, USA, July–Aug 2017
14. Fu, X., Zhuang, P., Huang, Y., Liao, Y., Zhang, X.P., Ding, X.: A retinex-based enhancing approach for single underwater image. In: IEEE International Conference on Image Processing, pp. 4572–4576 (2014)
15. Li, C.-Y., Guo, J.-C., Guo, C., Cong, R.-M., Gong, J.: A hybrid method for underwater image correction. Pattern Recognit. Lett. **94**, 62–67 (2017)
16. Ancuti, C.O., Ancuti, C., De Vleeschouwer, C., Bekaert, P.: Color balance and fusion for underwater image enhancement. IEEE Trans. Image Process. **27**(1), 379–393 (2018)
17. Buchsbaum, G.: A spatial processor model for object colour perception. J. Frankl. Inst. **310**, 1–26 (1980)
18. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1956–1963 (2009)
19. Carlevaris-Bianco, N., Mohan, A., Eustice, R.M.: Initial results in underwater single image dehazing. In: Proceedings of the IEEE conference on OCEANS 2010, Seattle, WA, USA, Sept 2010
20. Duarte, A., Codevilla, F., Gaya, J.D.O., Botelho, S.S.C.: A Dataset to Evaluate Underwater Image Restoration Methods, in Proceedings of the IEEE conference on OCEANS 2016, Shanghai, China, April (2016)
21. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 1–14 (2004)

Technique of Face Recognition Based on PCA with Eigen-Face Approach

C. B. Tatepamulwar, V. P. Pawar, S. D. Khamitkar and H. S. Fadewar

Abstract PCA is utilized in the area of recognition of face, fingerprint, handprint, industrial robotics, and mobile robotics. In the face recognition, research shows that the success rate is not satisfactory for a variant of poses which have rotation gap of more than 30°. If there are lots of variations in lightning, expressions, and pose variation, then PCA results are not up to the mark in the existing algorithm. This problem is arising in mind. The objective of the present paper is to study and propose modified PCA and Eigen-face-based algorithm to improve result with the accuracy of face recognition. In this paper, we focus on the pose variations which have 30° range of pose in image.

Keywords PCA · Eigen-face · Euclidian distance

1 Introduction

Biometrics is the emerging area of computer engineering; it is the method of recognizing a person based on a physiological or behavioral characteristic. The existing several biometric systems are available; among these facial recognition are one of the most universal, collectable, and accessible systems. Face recognition is the identification of a person from an image trained database. There may be variations in faces due to pose variation, age, hairstyle, etc. As per the vast literature review on the topic, there are two classifications of the existing face recognition techniques [1].

C. B. Tatepamulwar · S. D. Khamitkar · H. S. Fadewar (✉)
School of Computational Sciences, SRTM University, Nanded, India
e-mail: fadewar_hsf@yahoo.com

V. P. Pawar
WALMI, Kanchanwadi, Aurangabad, India

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_92

1.1 Holistic Approach

Input data of the face region for face detection is followed in holistic approach method. This holistic approach method is mostly used for Eigen-faces, fisher-faces, support vector machines, nearest feature lines (NFLs), and independent component analysis approaches [2]. Eigen-face method is based on principal component analysis (PCA) techniques which can be used to simplify a dataset into lower dimension.

1.2 Feature-Based Approach

Features on face such as nose, and then eyes are segmented and then used as input data for structural classifier which is used in Feature-based face recognition approach.

2 PCA (Principal Components Analysis) Technique

In 2011, Zhang Haiyang in his paper Face recognition based on DCT and PCA say that research in the area of facial biometric shows that PCA is an effective method for face recognition. In PCA technique, face images are converted into a set of eigenvectors and these eigenvectors is called eigen-faces [3].

2.1 Eigen-Face Approach

In Eigen-Face method, each feature or component is represented as an eigenvector of human faces. These eigenvectors do not correspond to the physical entities at the face, e.g., eye, nose, mouth, etc. In the process of recognizing human faces, each face is projected into a set of Eigen-face features. The eigenvectors corresponding to these features are weighted and the sum of these weights is a representation of a given face. Eigen-faces model is based on two-dimensional information only and each image can be represented as a matrix. A matrix has a set of eigenvectors that represent the principal components of the matrix [4]. Eigen-faces are the eigenvectors of the covariance matrix of all faces. Most of the research on Eigen-faces is concerned with frontal pose image and on rotating the image about the image center for pose variation. A collection of different face images of various people of different pose angles of a same person is needed to estimate the actual face. The specific necessary information of a face image needs to be extracted. This extracted information is compared with previously already defined database of face images [2].

The inspirations of Eigen-faces are given below:

- Capturing the statistical variation between face images is the correct way to extract the information. The appropriate features of face which is not necessarily associated with human facial attributes just like lips, nose, and eyes.
- Another motivation is that the system should efficiently represent face images. With the help of small number of parameters, we can represent the face image.

2.2 Generating Eigen-Faces

Basic requirement for using PCA-based Eigen-face approach is to find the Principal Components. Eigen-faces are nothing but the set of face images which are calculated using covariance matrix that contains eigenvectors. When the location points of each image shows corresponding eigenvector, and then it is convenient to represent those eigenvector of face [5, 6].

Let us assume that the particular face image is represented in the form of (p, q). This is nothing but the two-dimensional representation of image in form of matrix [6]. Let us suppose the dimension of image matrix is N-by-N array. It can also be presented in form of dimensions of vector as N^2.

2.3 Eigenvectors and Eigen Values

Nonzero linear operator vector elements in the linear algebra are called as Eigenvectors. Whenever any operations are implemented on these elements, multipliers of scalar value of these eigenvectors are generated [7]. This scalar value called as Eigen value is usually denoted as λ. These Eigen values are associated with eigenvector denoted by X and vector function is denoted by A

$$AX = \lambda X, \tag{1}$$

Eigenvector And Eigen value Calculation

Equation (1) can calculate value and form the equation as below. Identity matrix in the equation is denoted by

$$(A - \lambda I)X = 0, \tag{2}$$

Basic mathematical approach is going to form with the help of the above equation. This is called as equations of homogeneous system. Fundamental linear equation is given as below. The det variable in this equation is defined as determinant.

$$\det A - \lambda I = 0, \tag{3}$$

Fig. 1 Formation vectors of an image

After evaluations, n degree of polynomial is generated. This form of equation is called as the characteristic equation of A. If generated polynomial is related with characteristic equation, then this is known as characteristic polynomial. This can be always in form of n degree [3, 8].

$$\mathrm{A\,Xi} = \lambda\,\mathrm{Xi}, \text{where}\ \mathrm{i} = 1, 2, 3 \ldots, \mathrm{n} \tag{4}$$

There are n related or corresponding linear independent Eigenvectors present, if the Eigen values are all different.

2.4 Representing Face Image

A face image can be seen as a vector. Construction of vector of an image is formed by a simple merging of all lines from an image and is placed side by side with the other, as seen in Fig. 1.

Input test face image used for processing is 150×120 two dimension vector.

Training Set
Generation of training set is the prime step for representing the face image. For that purpose, let us assume that $N \times N$ denotes m images which are saved in training set. Afterwards, assume that Γ1, Γ2, Γ3,…, ΓM is the face image training set. After generating the training set for images, $N \times N$ matrix stores the face image feature vector value [4].

Mean of Face Image
Mean of face image can be found using the following formula.

$$\Psi = (1/M)\sum_{i=1}^{M} \Gamma_i \tag{5}$$

$$\Psi = (\Gamma_1 + \Gamma_2 + \Gamma_3 + \cdots + \Gamma_M)/M \tag{6}$$

Each face differs from the average by $\Phi_i = \Gamma_i - \Psi$ which is called mean centered mage [9].

Mean Subtracted Image

$$\varphi = \Gamma - \Psi \tag{7}$$

To differentiate mean image and training image from one another, the difference is being calculated. The difference between mean and training image results in mean subtracted image [7].

Covariance Matrix
Values of all mean subtracted training images are stored in one matrix and are known as covariance matrix. A covariance matrix is constructed using Eq. 8.

$$\mathrm{C = A\,.\,A^T, where\ A = [\Phi 1,\ \Phi 2,\ \Phi 3 \ldots \Phi 4]\ of\ size\ N^2 \times N^2} \tag{8}$$

Size of covariance matrix will be N2 × N2 (4 × 4 in this case). Eigenvectors corresponding to this covariance matrix is calculated, but that will be a complicated task [6, 10]. Therefore, just calculate ATA which would be a 2 × 2 matrix in this case. Hence, the size of this matrix is M × M. Let us assume the eigenvectors υi of AT A like that

$$\mathrm{A^T A\, X_i = \lambda_i\, X_i} \tag{9}$$

The Eigenvectors υi of AT A are X1 and X2. Multiplying the above equation with A forms other equation as below

$$\mathrm{A\,A^T A\, Xi \ = A\,\lambda_i\, X_i} \tag{10}$$

$$\mathrm{A\,A^T (A\,X_i) = \lambda_i (A\,X_i)} \tag{11}$$

3 Classification of Face Image Using Eigen-Face

The Eigenvectors of the covariance matrix $\mathrm{AA^T}$ are $\mathrm{AX_i}$ which is denoted by $\mathrm{U_i}$. The $\mathrm{U_i}$ can represent the Eigenvector of the covariance matrix $\mathrm{AA^T}$. This images looks ghost faces, hence it is known as Eigen-faces. Discard those faces which eigen value are zero for reducing the eigen-face space [5, 11].

3.1 Projecting Face Image

In this process, space of Eigen-faces is projected by training images. After being projected, Eigenvectors weight value is calculated. Each image in the Eigen-face space is denoted by computed weight value. Weight value is calculated by multiplying each image with corresponding Eigenvectors.

Projection of face image onto face space is shown by the formula given as below.

$$\Omega_K = U^T(\Gamma_k - \Psi); k = 1, 2, \ldots, M \tag{12}$$

And $(\Gamma_k - \Psi)$ is the mean centered image.

4 Recognition Process Steps

4.1 Finding Euclidean Distance

Widely used distance measure in the linear algebra is the Euclidean distance. The distance between two dots those are present in Euclidean space is called as Euclidean distance. These two dots are connected by a straight line, it is also known as Euclidean metric [4, 5]. After estimating the distance, Euclidean space transforms into metric space. The following are important thing that needs to consider for recognition process [6, 10].

1. The first aspect is to compute distance between the test image and each of the images in database.
2. For getting the result of most similar person's face, select the image which is most resemble to the new one.
3. To recognize the face image, decide the specified threshold value. Face image is recognized, if the distance of image is above value of threshold; if not then categorize face as unmatched face [9].

Euclidean distance between related feature vector values is computed to get the resemblance between pair of face image. This task can be performed after completion of projection of face image [5, 6]. A distance of Ω, i.e., feature vector to each face is called Euclidean distance and is defined using the formula given below. Vector of face class is denoted by k

$$\varepsilon_k^2 = \|\Omega - \Omega_k\|^2; k = 1, \ldots, M \tag{13}$$

When value of εk found the below-defined threshold value θc and if found to be minimum, then face belongs to the class k, if not then face is categorized as unknown face. A threshold distance can be defined for the face class with particular specified maximum value. θc is the half largest distance between any two face images [7].

$$\theta_c = 1/2\, max \left\| \Omega_j - \Omega_k \right\|; j, k = 1, \ldots, M \tag{14}$$

To find out the distance, first we need to form reconstructed image from Eigen-face Γ^f using Eq. 15

$$\Gamma^f = U * \Omega + \Psi \tag{15}$$

Next is to compute the distance ε between test image Γ and reconstructed image.

$$\varepsilon^2 = \|\Gamma - \Gamma f\| \tag{16}$$

If $\varepsilon \geq \theta_c$, this is the case, then the test image is not in space and not recognized.

If $\varepsilon < \theta_c$ and $\varepsilon_k \geq \theta$ in this case for all k, then the test image is a face image but it is still recognized as unknown face, i.e., false acceptance.

If $\varepsilon < \theta_c$ and $\varepsilon_k < \theta$ for all k, then the test images is recognized.

5 Existing PCA-Based Eigen-Face Approach Face Recognition Algorithm

1. Steps of existing face recognition algorithm is given below.
2. Read input image
3. Prepare training face database
4. Normalize images from training set
5. Select dataset
6. Find mean face factor
7. Subtract mean face factor from original faces
8. Compute covariance matrix
9. Find Eigen value and eigenvector for covariance matrix
10. Generates Eigen-Faces
11. Create reduced Eigen-face space
12. Compute Euclidean distance between image and the Eigen-faces.
13. Find the minimum Euclidean distance and perform recognition.

5.1 Limitations in Existing PCA-Based Face Recognition Algorithm

Existing algorithm uses many operations for face recognition process. It is the—-consuming task of existing algorithm for generation of training set. Generation of training set, normalization of training set, and then selecting the dataset are very lengthy process for performing further recognition operation. The accuracy of existing face recognition algorithm for pose variation of face poses is less. In this paper, we are focused on the face image in large pose variations [12].

6 Proposed PCA-Based Eigen-Face Approach Face Recognition Algorithm

In this study, face recognition is performed under the Eigen-face method which is the appearance-based technology. Once face detection got successful, then the system should perform face recognition of the particular selected image. The following procedure shows face recognition step by step using MATLAB functions. In this way, face recognition systems algorithm is implemented as shown in Fig. 2.

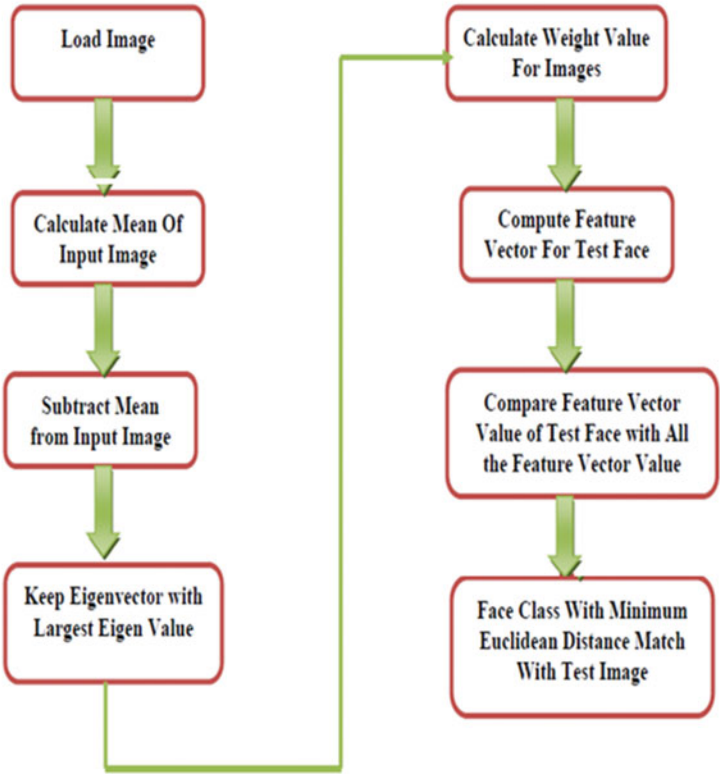


Fig. 2 The proposed face recognition algorithm

7 Image Data Base for Input Test Images and Trained Face Image Database for Recognition

In our database, it is classified into input test images and output trained database images. In input test image, database contains 20 subjects. Each subject contains seven profile images, one front profile image, three left images, and three right images from pose variation angel +90° to −90° with gap of 30°. So, total number of images is $20 \times 7 = 140$ input test images. In trained image, database contains 20 subjects of front profile or average images. Images of the train database are considered for face recognition. Images of the train database have numbering from *Person 1 to Person 20*, respectively, used for reference purpose. All sample face images (150 × 120 pixels) are stored in the database in Portable Network Graphics (png) file format [6, 10] (Fig. 3).

8 Results and Analysis

The face recognition method is implemented with MATLAB 2012. Testing is performed with our own database, i.e., SRTMUNFD and acquired images as well. In face recognition step, input test image is acquired. This input test image will be compared with the all face images stored in train database. Initially, Eigen-faces for all face images stored into train database are calculated. PCA algorithm is used to calculate the Eigen-face. After finding Eigen-faces for trained face images, Eigen-face for input test image is also calculated. After that, Euclidean distance is calculated from input test face to all trained faces [11]. Face image with the minimum Euclidean distance is selected as recognized face (Figs. 4 and 5).

From the above graph, minimum Euclidean distance between input test image and all trained database image is 4210.9832, so the recognized image is person 13 from trained database.

For evaluating interpretation of our face recognition system, we have tested our proposed algorithm on our own database in natural environment with varying pose. This processing gives us good performance over images with different poses. In our study, we have total 140 images out of which 20 are frontal view images and the remaining 120 are profile view images. We resized all images in one fixed image resolution and then carried out face recognition.

In the proposed algorithm, image list is generated. For carrying out recognition of faces, each image in database is compared with Eigen-faces. After comparison, the image which has minimum Euclidean distance is supposed to be recognized image, if computed distance is higher or maximum, then image considered as not recognized. When we take test image as an input image, then it returns image which is frontal display original one. Generally, matching criteria is depending on minimum Euclidean distance. But we are snidely changes in this criterion. In this paper stored value of feature vector in .mat file in Matlab. Feature vector value which is called as

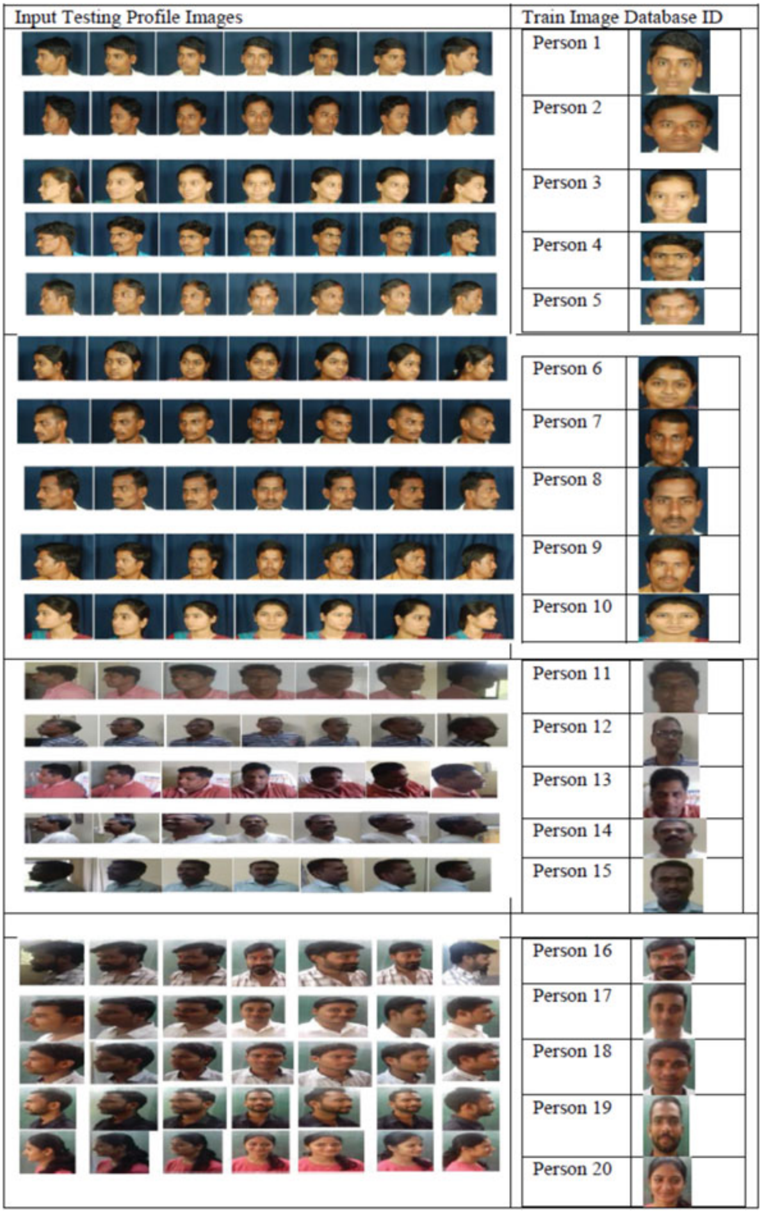


Fig. 3 Input test profile images and training front profile images database

weight factor of input image is equated with all the feature vector value in .mat file. Then, it computes Euclidean distance between feature vector value of input image and feature vector value of remaining image. Finally, face with minimum Euclidean

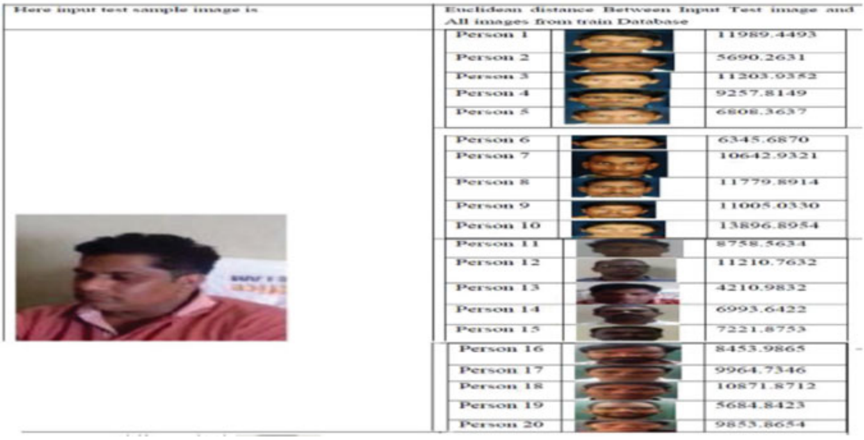


Fig. 4 Experiment result for person 13

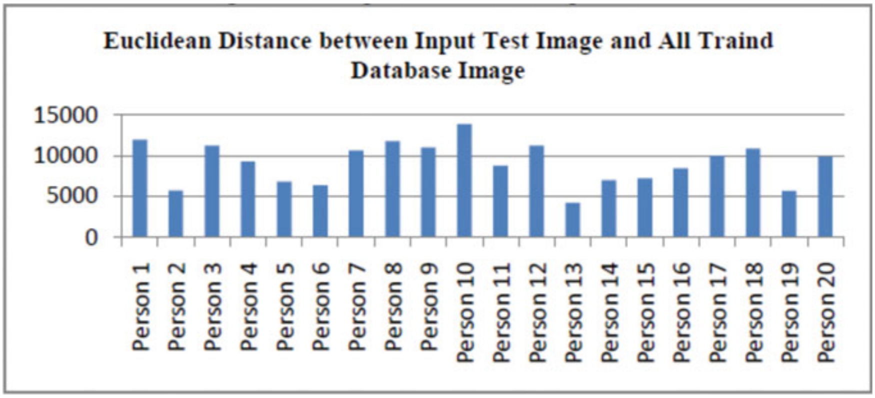


Fig. 5 Euclidean distance measure for person 13

distance found to be match. Total number of images in the database including profile and frontal is given. The following table is presented that shows matching percentage area of each pose of face image.

The future applications will be dominated by IoET [9]. In view of this, the present work will be extended as an IoET-enabled system with more intelligence and sophistication.

9 Conclusion

In this paper, face recognition is performed with the help of PCA-based method. PCA is a feature extraction method which uses Eigen-face for matching purposes. When test image is read, it forms its mean image and generates the Eigen-faces. The detailed pictorial implementation is given for both frontal view and profile view face image. The proposed method recognizes face image in both the cases, i.e., frontal view and posed view. The system should recognize the persons face image with pose variations, and the proposed method solves this problem which usually happens in real-time face recognition system.

Declaration The work reported in this chapter is approved by the ethical approval committee of School of Computational Science, SRTMU, Nanded (MS)-India. The committee consists of Dr. G. V. Chowdhary (Chairman), Dr. S. D. Khamitkar, Dr. H. S. Fadewar, Mr. M. D. Wangikar. Further, the subjects under test had given their written consent for the experiments and publication of this work.

References

1. Zhujie, Y.L.Y.: Face recognition with Eigen faces. In: Proceeding of IEEE International Conference on Industrial Technol, USA, pp. 434–438 (2011)
2. Patil, M., Iyer, B., Arya, R.: Performance evaluation of PCA and ICA algorithm for facial expression recognition application. In: Proceedings of Fifth International Conference on Soft Computing for Problem Solving, pp. 965–976 (2016)
3. Terence, S., Rahul, S., Mathew, M., Shumeet, B.: Memory based face recognition for visitor identification. In: Proceedings of International Conference on Automatic Face and Gesture Recognition, Pitresburg (2000)
4. Turk, Pentland: Face Recognition Using Eigenfaces, IEEE CH2983-5/91, pp. 586–591
5. Wang, L., Tan, T., Hu, W.: Face tracking using motion-guided dynamic template matching. In: 5th Asian Conference on Computer Vision, 23–25 January 2002, Melbourne, Australia, pp. 1–6
6. Tian, Y., Kanade, T.: Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell. **23**(2) (2011)
7. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: International Conference on Neural Information Processing Systems (NIPS), 2008
8. Torre, F.D.L., Black, M.J.: Face recognition based on eigen faces. In: IEEE International Conference on Computer Vision (ICCV '2001), Vancouver, Canada, July 2010
9. Deshpande, P., Iyer, B.: Research directions in the internet of every things (IoET). In: International Conference on Computing, Communication and Automation (ICCCA), pp. 1353–1357 (2017)
10. Pentland, A., Moghaddam, B., Starner, T.: View-based and modular eigenspaces for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 21–23 June 2010, Washington, USA, pp. 84–91
11. Haiyang, Z.: Face Recognition Based on DCT and PCA. In: Lecture Notes in Electrical Engineering (2011)
12. Phiasai, T., Arunrungrusmi, S., Chamnongthai, K.: Face recognition system with PCA and moment invariant method. In: Procceeding of IEEE Inernational Symposium on Circuits System, pp. 165–168 (2001)

Analysis of Face Recognition Algorithms for Uncontrolled Environments

Siddheshwar S. Gangonda, Prashant P. Patavardhan and Kailash J. Karande

Abstract Face recognition is a challenging problem in biometric systems, which has received a lot of attention in the last two decades as it has numerous applications in computer vision and pattern recognition. There is remarkable progress in the face recognition systems under controlled conditions, but they degrade for uncontrolled conditions like pose, illumination, expression, and occlusion etc. In this paper, we discussed different algorithms like PCA, DCT, LDA, ANN, ICA, HMM, and Wavelet with its pros and cons. The different face database used for face recognition is discussed. It also discusses various challenges and possible future directions for face recognition task.

Keywords Face recognition · Artificial neural network (ANN) · Discrete cosine transform (DCT) · Hidden Markov model (HMM)

1 Introduction

Face recognition in uncontrolled conditions is still a challenging problem. It has been an interesting and demanding research area. The numerous algorithms and modeling techniques have been developed. The most widely used algorithms for face recognition are PCA, DCT, LDA, ANN, ICA, HMM, Wavelet. Face recognition has an important advantage over many biometric technologies like it is a nonintrusive, noninvasive, and easy to use method [1]. Face recognition is a challenging problem in biometric systems, which has received a lot of attention in the last two decades as

S. S. Gangonda (✉) · K. J. Karande
SKN Sinhgad COE, Pandharpur, Maharashtra, India
e-mail: sgangonda@gmail.com

K. J. Karande
e-mail: kailashkarande@yahoo.co.in

P. P. Patavardhan
KLS Gogte Institute of Technology, Belagavi, Karnataka, India
e-mail: prashantgemini73@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_93

it has numerous applications in different fields. Hence, it is essential to develop an automated system for personal identification which is needed for security systems, ATM access and crime detection. The changes in expression, lighting conditions, pose, and occlusion makes face recognition difficult and a challenging task. The reliability and robustness are very important for these face recognition applications particularly in security systems.

2 Challenges in Face Recognition

The various challenges in face recognition occur in real conditions which degrades the recognition performance. The main challenges of face recognition are as follows (Fig. 1) [2].

- **Facial expression variations**: The different facial expressions like anger, laughing, yawing, smile, etc., changes the normal (the neutral expression) look and structure of a face which affects the recognition result.
- **Illumination variations**: The light from different directions produces a dramatic change in facial look. The changing direction and energy distribution of the surrounding illumination along with the 3D structure of the human face produces greater differences in the shadows and shading on the face.

Fig. 1 Various challenges of face recognition system [3]

- **Pose variations**: Pose is another challenge in the recognition of faces as the alignment of different 2D face images having unlike poses are a difficult task.
- **Occlusion**: It is a challenge in the recognition of faces which is caused when an external object (other than the original face area) is present in between the face surface and the camera. The common occlusions are the accessories like sunglasses, hat, scarf, phone, and hand on a face. The other sources of occlusions are cosmetics and face behind fence.
- **Aging**: The human face is ever changing. As the age of a person increases, the look of a person also goes on changing which affects the performance of the face recognition system.
- **Identify similar faces**: The two different persons may have similar look like the twins which is impossible even for a human to identify them. Hence, it affects the performance of the face recognition system.

2.1 Face Databases

The different face databases have been built for the analysis of face images when dealing with a single or a combination of these changes. The different types of face image databases are given in the Table 1.

The image size, image type, and the other specifications describes about the complexity of face database which in turn shows the robustness of different algorithms of face recognition. The different face databases are created to evaluate the effect of changes on the several types of conditions of an image [4]. AR Faces, FERET, CMU-PIE, Asian and Indian face database are the most widely used 2D face image databases. Each database provides a platform to access the particular challenges of uncontrolled conditions. For example, CMU-PIE is used for more illumination and poses changes. FERET gives a good testing platform for large probe and gallery sets. AR faces gives the natural occluded face images. Asian face database consists of 2D

Table 1 Different types of face image databases [4, 5]

Name	Image type RGB/gray scale	Image size	Types of conditions
FERET	Gray RGB	256×384	i, e, p, I/O, t
The yale face B	Grayscale	640×480	i, p
AR Faces	RGB	576×768	i, o, t, e
CMU-PIE	RGB	640×486	i, e, p
The yale face	Grayscale	320×243	i, e
Asian face database	RGB	640×480	p, e, i, o
Indian face database	RGB	640×480	e, p

The different Image changes are shown by p: pose, o: occlusion, i: illumination, e: expression, t: time delay, I/O: indoor/outdoor conditions face images of female and male with pose, illumination, expression occlusion, and expressions. The Indian face database comprises of face images with variation in expression and poses [2].

3 Face Recognition Algorithms

3.1 Introduction

The various algorithms that are used for the recognition of face images are Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), Discrete Cosine Transform (DCT), Artificial Neural Network (ANN), Independent Component Analysis (ICA), Hidden Markov Model (HMM), and Wavelet, etc. In this, we review these algorithms briefly.

3.1.1 Principal Component Analysis

PCA is most widely used method for face recognition which reduces the dimensionality of the images. Hence, it recognizes the images quickly without disturbances. PCA is also known as Karhunen–Loeve method [6] which has poor discriminating power. The recognition performance of PCA reduces under changing illumination and poses [7]. It uses the scale normalization in the process of recognition of face images. The advantage of this method compared to other face recognition systems is its speed, simplicity, and insensitivity to minor variations on the face [8].

3.1.2 Discrete Cosine Transform

DCT has the best energy compaction feature which helps in reduction of the dimensionality of the face images. It is useful in the data compression. The first coefficients give the major useful information from the original data. The DCT method computed efficiently which makes it useful in recognition of the face images. This method has precise and robust recognition of the face images which uses the techniques of normalization. The affine transformation is incorporated in it which autocorrects the changes in faces.

3.1.3 Independent Component Analysis

ICA is a method which is iterative and it has a powerful data representation than PCA. Hence, it has better recognition rate compared to PCA. It does not offer ordering of source vector. ICA transforms the data as linear alliances of statistically autonomous data points [9].

3.1.4 Wavelet

Wavelet is the prominent multi-resolution method used for the recognition of face images. Due to the best time–frequency localization feature, it gives significant performance in the face recognition task. It works suitably well under large illumination and pose changes than DCT and PCA techniques.

3.1.5 Linear Discriminant Analysis

It is another powerful method used for the recognition of face images. It has an effective representation which linearly transforms the main data space into a low-dimensional feature space. It is useful in reducing the dimensionality of the image due to which it gives the best performance in the recognition task. LDA method performs better than PCA. The limitation of this method is that it encounters the small sample problem. It also fails when scatter matrices are singular called as singularity problem.

3.1.6 Artificial Neural Network

ANN is a technique used for the recognition of face images which has single-layer adaptive network. Due to the simplicity and ability in pattern matching, Multi-Layer Perceptron (MLP) is used along with feed-forward learning techniques [6, 10].

3.1.7 Hidden Markov Model

It is one of the prominent methods used for the recognition of face images having compatibility with the uncontrolled conditions like expression, illumination and orientation. HMM is a method which has a set of statistical models expressing the special features of the signals. Due to the presence of one-dimensional data in it, HMM has excellent performance in character recognition and speech recognition. The purpose is to get the hidden parameters out of the observable parameters.

3.1.8 Face Recognition Using Combined Approaches

The robust face recognition system can be designed by combining the different face recognition algorithms, approaches which are invariant to pose, illumination and occlusion. The Face Analysis for Commercial Entities (FACE), a novel framework for face analysis is proposed in which the robustness comes from Normalization strategies to address pose and illumination variations. The robust face recognition in uncontrolled illumination and occlusion by Structured Sparsity is performed by integrating Structured Sparsity Representation with robust features of discriminative feature Weber Local Descriptor.

3.2 Analysis of Face Recognition Algorithms

The analysis of most popular algorithms of face recognition (PCA, ICA, and LDA) in terms of memory usage, recognition rate, and data representation are given in Table 2.

The PCA method consumes high memory for the storage purpose and it has powerful representation of the data. It has recognition rate better than Eigen and Fisherface. The ICA method consumes moderate memory for the storage purpose, and it has the representation of the data powerful than PCA. It has recognition rate better than PCA. The LDA method consumes low memory for the storage purpose, and it has the strong representation of the data. It has recognition rate better than PCA. The pros and cons of the different algorithms used for recognition of the face images (PCA, ICA, and LDA) are discussed in Table 3.

Table 2 Analysis of face recognition algorithms [9]

Technique	Memory usage	Recognition rate	Data representation
PCA	High	Better than eigen and fisherface	Powerful data representation
ICA	Moderate	Better than PCA	Powerful than PCA
LDA	Low or efficient	Better than PCA	Strong data representation

Table 3 Pros and cons of face recognition algorithms [9, 11]

Technique	Pros	Cons
PCA	• Reduces dimensionality of image • Simple, fast, and robust • Image without disturbance are Recognized fast	• Poor discriminating power • Reduces recognition under changing illumination and pose • Requires scale normalization
ICA	• Powerful data representation • Better recognition rate compared to PCA	• It is iterative and coverage difficult • It does not offer ordering of source vector
LDA	• Face recognize of the image without disturbance • Overcome the limitation of PCA	• Encounters small sample problem • Fails when scatter matrices are singular called as singularity problem

3.3 Future Research Directions

The major new research needed to be focused on minimizing the various challenges in face recognition techniques. The problems of face recognition can be resolved by extracting the static features from the face images or by designing the machine learning modeling techniques. The robust face recognition system can be designed by combining the different face recognition algorithms.

Another research direction would be to reduce the computational complexity of the various face recognition algorithms so that they could be used in practical world commercial products [5].

4 Conclusion

Face recognition is a challenging problem in biometric systems, which has received a lot of attention in the last two decades as it has numerous applications in different fields. In this paper, we discussed different algorithms like PCA, DCT, LDA, ANN, ICA, HMM, and Wavelet with its pros and cons. The analysis of different algorithms used for recognition of the face images like PCA, ICA, and LDA in terms of memory usage, recognition rate, and data representation are done. The different face database used for face recognition is discussed. It also discusses various challenges and possible future directions for face recognition task.

References

1. Li, S.Z., Jain, A.K.: Handbook of Face Recognition. Springer, London Dordrecht, Heidelberg, NewYork. ISBN 978-0-85729-931-4
2. Karande, K.J., Talbar, S.N.: Independent Component Analysis of Edge Information for Face Recognition. Springer Briefs in Applied Sciences and Technology (2014). ISSN: 2191-530X
3. Bhele, S.G., Mankar, V.H.: A review paper on face recognition techniques. Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) **1**(8) (2012). ISSN: 2278-1323
4. Bakhshi, Y., Kaur, S., Verma, P.: A study based on various face recognition algorithms. Int. J. Comput. Appl. **129**(13), 0975–8887 (2015)
5. Ohlyan, S., Sangwan, S., Ahuja, T.: A survey on various problems & chal- lenges in face recognition. Int. J. Eng. Technol. (IJERT) **2**(6), (2013) ISSN: 2278-0181
6. Kakade, S.D.: A review paper on face recognition techniques. Int. J. Res. Eng. Appl. Manag. (IJREAM) **02**(02) (2016). ISSN: 2494-9150
7. Sharif, M., Mohsin, S., Javed, M.Y.: A survey: face recognition techniques. Res. J. Appl. Sci. Eng. Technol. **4**(23), 4979–4990 (2012). ISSN: 2040-7467
8. Singh, K.R., Zaveri, M.A., Raghuwanshi, M.M.: Illumination and pose invariant face recognition: a technical review. Int. J. Comput. Inf. Syst. Ind. Manag. Appl. (IJCISIM) **2**, 029–038 (2010). ISSN: 2150-7988
9. Aswathy. R.: A literature review on facial expression recognition techniques. IOSR J. Comput. Eng. (IOSR-JCE) **11**(1), 61–64 (2013)

10. Saini, R., Saini, A., Agarwal, D.: Analysis of different face recognition algorithms. Int. J. Eng. Res. Technol. (IJERT) **3**(11) (2014). ISSN: 2278-0181
11. Hassaballah, M., Aly, S.: Face recognition: challenges, achievements and future di- rections. Inst. Eng. Technol. (IET) (2014). ISSN: 1751-9632. https://doi.org/10.1049/iet-cvi.2014

Line Scratch Detection in Old Motion Picture

Mukkawar Vinayak and Jondhale Kalpana

Abstract Detecting line scratch in motion is a tedious job because it requires spatial as well as temporal features needs to be extracted. Scratches are caused by abrasion of film material as it passes through projection mechanism. One main problem in this is false detection. Spatial algorithm is used to detect scratch inside the frame using frame-wise scratch detection and temporal algorithm is used for filtering false detection. Preprocessing is required to get fine results. Experiment result shows the detection of line scratches in motion picture.

Keywords Line scratch · Detection · Spatial filtering · Temporal filtering

1 Introduction

The detection of line scratches in motion picture is a fundamental task in video restoration due to the presence of large amount of old material in film archive. To find the defect in the image manually is tedious and time-consuming job. Some of the common defects in the films include dirt/dust, flicker, blotches and line scratch. In this paper, we deal with last defect that is line scratch, usually caused by abrasion to the physical film due to the projection mechanism. Line scratches are thin either dark or bright lines which are mostly straight and vertical because of projector motion in vertical direction. The information of the physical origins of line scratches may be found at [1]. The line scratches represents characteristic of temporal persistence, indicating that the scratches remain in same position for several frames. However, this characteristic is variable, for example consider the image having scratches not

M. Vinayak (✉)
Department of IT, MGM's COE, Nanded, Maharashtra, India
e-mail: mukkawar_vv@mgmcen.ac.in

J. Kalpana
Department of ETC, MGM's COE, Nanded, Maharashtra, India
e-mail: jondhale_kc@mgmcen.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_94

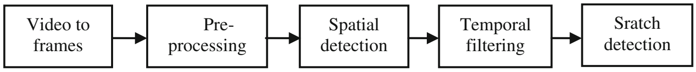


Fig. 1 Flowchart of scratch detection

necessarily completely straight and vertical and their shape may vary from frame to frame.

We design the line scratch algorithm in which first, we detect the position of pixel representing line scratch, and in this detection, we get much false detection, so to eliminate false detection, we use motion coherency. Some of the algorithm like [2] uses energy regularity to find out line scratch detection.

2 Proposed Algorithm

Figure 1 represents the flow of proposed method which consists of video-to-frame conversion, preprocessing the image to avoid noise as well as false detection, spatial detection is used to detect significant scratch segment by considering the current frame, temporal scratch detection is used to eliminate false detection using temporal information contained in the image sequence.

2.1 Spatial Detection

First convert the video into frame and apply Gaussian filter of size 3×3 with standard deviation of one pixel on it. Gaussian-filtered image is then passed to spatial detection. Spatial detection consists of two steps,

2.1.1 Pixel Wise Scratch Point Detection

This step detects potential scratch points. Some other method to detect scratch is found in [3]. This step consists of threshold on the difference between grayscale image and the median-filtered version of the same image [4]. This step is nearly same as explained by Kokaram in [5]. So, pixel-wise scratch point detection criteria can be given as

$I_g(x, y)$: Gaussian filtered gray level image
$I_m(x, y)$: Median value over neibhouring pixel(x,y)
$I_l(x, y)$: Left horizontal average
$I_r(x, y)$: Right horizontal average
S_{med} : Gray level threshold having value 3
S_{avg} : Graylevel threshold having value 20

$$c1(x, y) : |I_g(x, y) - I_m(x, y)| >= S_{med} \tag{1}$$

$$c_2(x, y) : |I_l(x, y) - I_r(x, y)| <= S_{avg} \tag{2}$$

$$I_b(x, y) = \begin{cases} 1 \text{ If c1(x, y) and c2(x, y)} \\ 0 \text{ else} \end{cases}$$

where $I_b(x, y)$ is binary detected image and this detection may produce many false detection. Figure 2 illustrates the binary detected image in which the upper image is original image and lower is its binary detected image. The white pixels in binary image are detected pixels and black pixels are not. As we see, there are many false detections therefore further grouping and validation step is needed in order to determine significant scratch segment.

2.1.2 Scratch Point Grouping and Validation

A robust method is needed to group the detected scratch point into scratch segments because of false detection due to noise as shown in Fig. 2. So, we use more a sophisticated method known as contrario method, which is used for alignment detection by Desolneux [6].

The contrario line segment detection considers the basis elements to be grouped are pixels and segment are detected as group of pixels whose gradients are perpendicular to given direction. Given a line segment made up of l pixels, a variable x_i is associated to each pixel. The variable x_i is 1 if pixel aligned with segment else it is 0. Pixels having gradient direction perpendicular to segment orientation with some angular precision $p\pi$ radians are said to be aligned pixel with $\in$ [0, 1]. Let $S = x_1 + x_2 + \cdots + x_l$ be the number of aligned pixels. Larger values of s are associated to more meaningful line segment. The aim of contrario method is to set the threshold value of l and p. The segment of length l having k_0 aligned pixels are meaningful when $B(p; k_0, l)$ distribution is small enough, a threshold must be assigned to this probability, so we take the total number of tested segment in account. For this, we consider the number of false detection (NFD) and is given as

$$NFD(l, k_0) = N_{test} B(p; k_0, l) \tag{3}$$

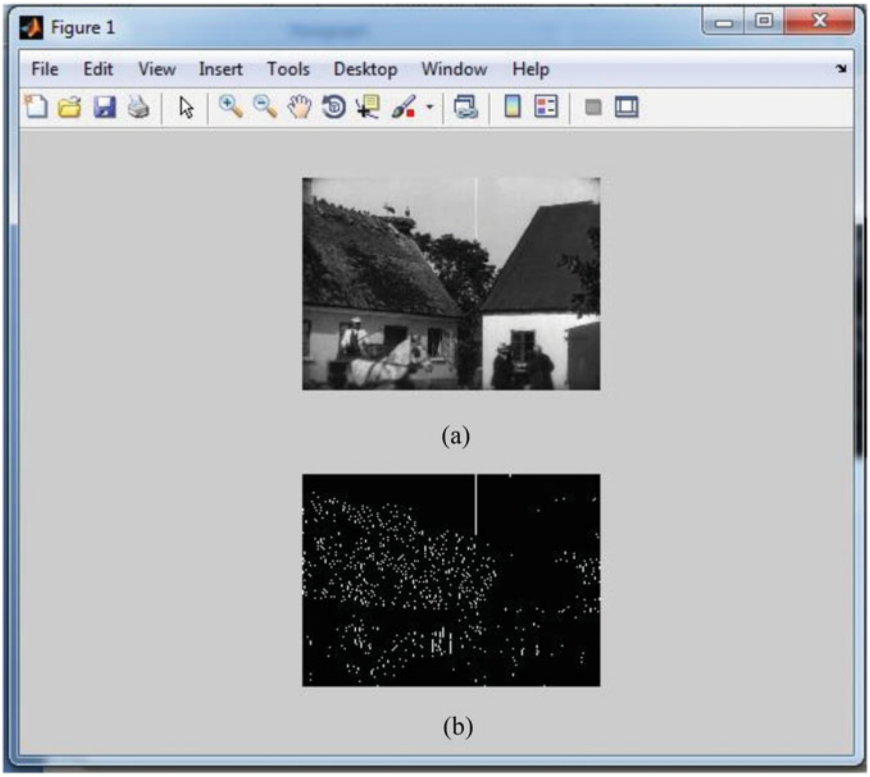


Fig. 2 **a** Gaussian-filtered image and **b** Binary detected image

where $N_{test} = M^2N^2$ total number of segment to be tested and M and N are linear dimensions of image. A segment is detected if $NFD(l, k_0) \le \varepsilon$ for some parameter ε [6].

If we use a binomial distribution, we get redundant segment because one meaningful segment contains other meaningful segment. So, number of false detection can be calculated using Poisson binomial distribution which can be given as

$$NFD(l, k_0) = N_{test}H(l, k_0) \tag{4}$$

where $H(l, k_0) := e^{-l(r \log \frac{r}{<p>} + (1-r)\log \frac{1-r}{1-<p>})}$ and

$$< p >= l^{-1} \sum p_i$$
$$r = \frac{k_0}{l}$$

Redundancy is much removed but some false detection's are there and are removed using temporal algorithm.

2.2 Temporal Filtering

Even though the line scratch detection algorithm detects the line scratches with good spatial precision and robust to noise, but it does not deal with the problem of false detection due to thin vertical structures that are part of frame. The pixel-wise detection criteria treats this thin structure as a line scratch, but originally they are part of image. So, we must have to develop new algorithm to detect these false detections. One way to detect true and false scratches is to use temporal information contained in image sequence. As we know that scratches are physical damage to film that is why their motion is completely independent to scene. So, motion coherent to scene represents the false detection.

First, we have to calculate the trajectories of false scratches which are calculated using algorithm from Odobez et al. [7] and the other approach may be found out at [8]. Using [7] algorithm, we estimate scenes global motion and realign the segment and are calculated using the following motion vectors:

$$u(q) = c_1 + a_1 x_q + a_2 y_q \tag{5}$$

$$v(q) = c_2 + a_3 x_q + a_4 y_q \tag{6}$$

where c_1 and c_2 are constant motion components and a_1 to a_4 are spatially varying components of motion.

3 Result

First, we will see some stepwise results and then we found F1 score. Figure 3 shows the detection of scratches in which upper image is single frame from video and lower image is scratch detection marked with green colour.

This section represents the quantitative result of our approach. We used three different criteria to calculate performance measure that are recall, precision, and F1 score. Recall can be defined as number of true detections divided by total number of true scratches that is percentage of line scratch detection. Precision is defined as number of true detections divided by total number of detections that is what percentage of detection was correct detection. The F1 score is nothing but the harmonic mean. The F1 score can be given as

$$F1 = 2\left(\frac{\mathrm{Re}call * \Pr ecision}{\mathrm{Re}call + \Pr ecision}\right) \tag{7}$$

Our two methods are compared with two other algorithms: the spatial method of Bruni [9] and temporal method of Gullu [7]. The comparison is shown in Table 1. It shows that in most of the cases, the result of our approach is better than other. The same thing is represented in graph as shown in Fig. 4. The X-axis represents the

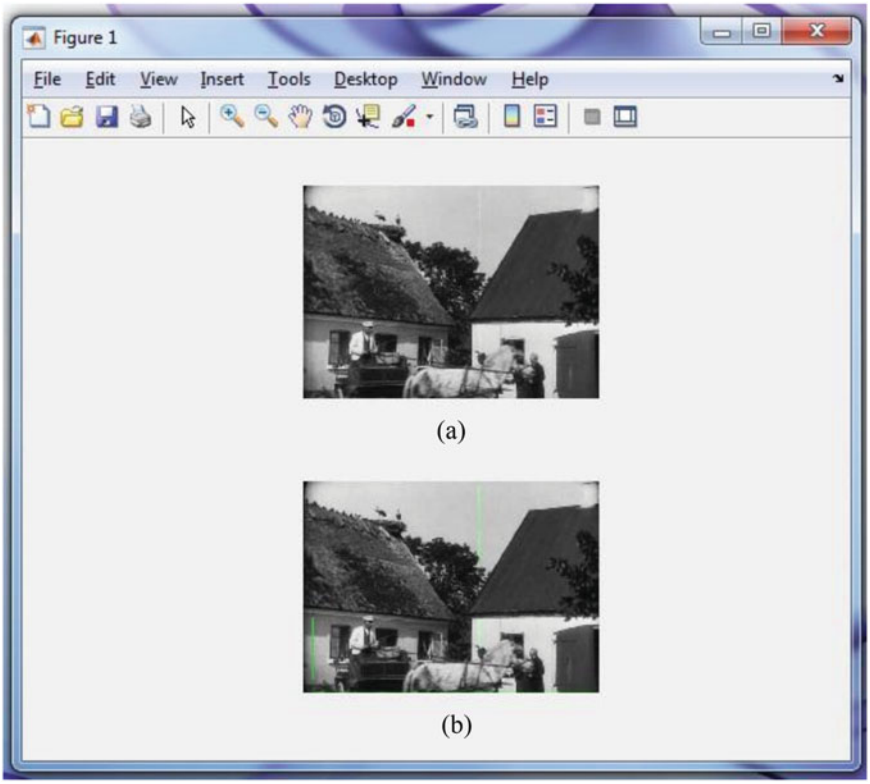


Fig. 3 **a** Frame in video and **b** Detected scratch in green color

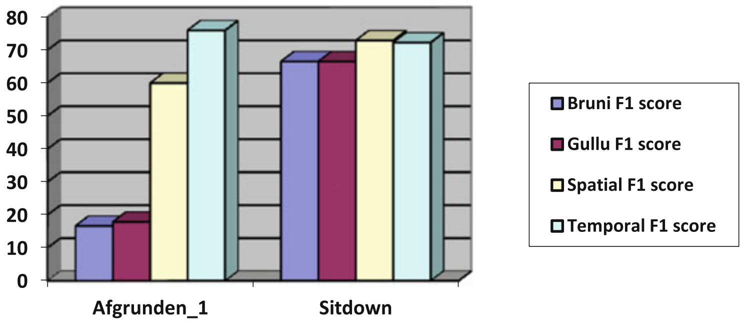


Fig. 4 Graphical representation of F1 score

videos on which proposed algorithms are applied and Y-axis represents the F1 score which is calculated using Eq. 7.

Table 1 Comparison of F1 score

Criteria	Algorithm	Video of work	
		Afgrunden_1	Sitdown
Recall	Bruni	75.11	80.93
	Gullu	68.75	80.93
	Spatial	86.66	73.72
	Temporal	86.66	72.63
Precision	Bruni	9.35	56.47
	Gullu	10.27	56.47
	Spatial	45.85	72.07
	Temporal	67.67	71.85
F1 score	Bruni	16.62	66.52
	Gullu	17.86	66.52
	Spatial	59.97	72.88
	Temporal	75.94	72.24

4 Conclusion

This paper represents one spatial algorithm to detect scratch point which uses the pixel-wise detection criteria and one temporal algorithm to minimize false detection which identifies scratches coherency with the scene.

There are several applications of proposed algorithm, and one can extend this work for video restoration by removing line scratch from digital films.

References

1. Physical Damage, National Film and Sound Archive. http://www.nfsa.gov.au/preservation/handbook/damage-films/physical-damage/
2. Kumar, R., Gupta, S., Venkatesh, K.S.: Detection of line scratch using energy regularity of spatio-temporal video cube. In: 2015 International Conference on Image Processing Theory, Tools and Applications (IPTA), Orleans, pp. 360–364 (2015)
3. Kim, K.-T., Kim, E.Y.: Film line scratch detection using neural network and morphological filter. In: The Proceeding IEEE Conference on Cybern Intelligent System, pp. 1007–1011, Sep 2008
4. Newson, A., Almansa, A., Gousseau, Y., Perez, P.: Robust automatic line scratch detection in films. IEEE Trans. Image Process. **23**(3) (2014)
5. Kokaram, A.: Detection and removal of line scratches in degraded motion picture sequences. Sig. Process. **1**(9), 5–8 (1996)
6. Desolneux, A., Moisan, L., Morel, J.M.: Meaningful alignments. Int. J. Comput. Vis. **40**, 7–23 (2000)
7. Gullu, M.K., Urhan, O., Erturk, S.: Scratch detection via temporal coherency analysis and removal using edge priority based interpolation. In: The Proceeding IEEE International Symposium on Circuits System, pp. 4591–4594, May 2006

8. Müller, S., Bühler, J., Weitbruch, S., Thebault, C.: Scratch detection supported by coherency analysis of motion vector fields. In: The Proceeding 16th IEEE International Conference on Image Processing, pp. 89–92, Nov 2009
9. Bruni, V., Vitulano, D.: A generalized model for scratch detection. IEEE Trans. Image Process. **13**, 44–50 (2004)

Underwater Image Enhancement by Rayleigh Stretching with Adaptive Scale Parameter and Energy Correction

Sonali Sankpal and Shraddha Deshpande

Abstract Attenuation of light in water causes degradation of underwater images. This attenuation is caused by water molecules, suspended particles, and dissolved chemical compounds in water. The attenuation includes scattering and absorption of light in water. Backward scattering and fading of color are two major sources of degradation of underwater images. This paper proposed a method of enhancement of underwater images by providing a solution for degradation because of backward scattering. The proposed method corrects the effect of backward scattering by enhancing contrast of the image by Rayleigh stretching of each color channel using maximum likelihood estimation of scale parameter. After contrast enhancement, loss of energy in the signal is corrected, that recovers information loss caused by contrast enhancement. The results of the proposed method are compared quantitatively with state-of-the-art methods by applying it to underwater dataset. Comparison is done with mean square error (MSE), Structural SIMilarity index (SSIM), and Average Information Entropy (AIE) quality metrics. It is seen that the proposed method in this paper produces best results when compared with state-of-the-art methods.

Keywords Rayleigh stretching · Contrast enhancement · Maximum likelihood estimation · Energy correction

1 Introduction

Underwater image processing has increased attention in the last few decades because of its enormous applications like, in unmanned vehicles, in sea installation and monitoring, study of underwater biodiversity, etc. Degradation of underwater images due

S. Sankpal (✉)
Department of Electronics & Telecommunication Engineering, PVPIT, Budhagaon, Sangli, Maharashtra, India
e-mail: sssankpal.etc@pvpitsangli.edu.in

S. Deshpande
Department of Electronics Engineering, WCE, Sangli, Maharashtra, India

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_95

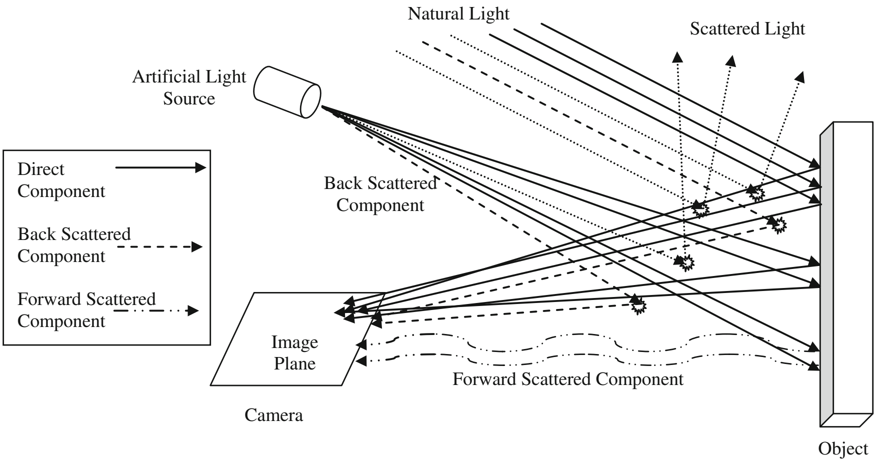


Fig. 1 Underwater imaging situation

to interaction of light with water, demands for processing of underwater images. This interaction is of two types, scattering and absorption, the scattering includes forward and backward scattering. Forward scattering is small angle scattering in forward direction which introduces resolution problem in image. Backward scattering is a scattering of light rays in backward direction in the image plane without reaching the object, which reduces the contrast of the image. Attenuation of light in water is function of a wavelength (or color). Among this, backscattering plays a major role in degradation of underwater images.

1.1 Background

According to Jaffe-McGlamery [1] image formation model, underwater image is formed by superposition of three components given in Eq. (1).

$$E_t = E_d + E_{bs} + E_{fs} \tag{1}$$

where E_t is total irradiance, E_d is direct component, E_{bs} is backscattered component, and E_{fs} is forward scattered component. Here, irradiance is radiant power per unit area. The direct component contributes to image without any attenuation. Forward scattered component is small angle scattered rays of light which contributes to image but provides strong resolution problem in the image, whereas backscattered component do not contribute to image and provides contrast degradation in the image. The underwater imaging situation along with all the above said components are shown in Fig. 1.

The attenuation of light in water is wavelength dependent [2], i.e., minimum for blue and maximum for red. Due to wavelength dependency of attenuation, colors in the light rays drop off as it travels in the water. Among all the colors in the light rays; blue travel longest, so underwater images are generally dominated by blue color. Light attenuation is also a function of a distance [2]. Equation (2) shows a decrease in intensity of unscattered light as a function of a distance for a particular wavelength λ.

$$I_r(\lambda) = I_0(\lambda)e^{-ar} \tag{2}$$

where I_r (λ) is intensity of unscattered light at distance "r" from camera for a particular wavelength λ. $I_0(\lambda)$ is initial intensity of unscattered light, "a" is attenuation coefficient. Typical values of attenuation coefficients for bay water, coastal water and deep ocean water are 0.33 m^{-1}, 0.2 m^{-1} and 0.05 m^{-1}, respectively, with visibility of 75, 125, and 500 m [3].

1.2 Motivation

Underwater applications required clear underwater images and it is achieved by image enhancement and restoration techniques. Restoration techniques need model of degradation, which requires model parameters like attenuation coefficient, turbidity, time, depth, etc. Enhancement techniques does not need these parameters and are much faster than restoration techniques. There are many researchers who attempted the problem of underwater image enhancement, but the area is in need of adaptive and fast image enhancement technique. This paper attempted underwater image enhancement technique to compensate for degradation of underwater images caused by backscattering of light in water. It is performed using contrast enhancement with histogram stretching with respect to Rayleigh distribution by estimating adaptive scale parameter. After contrast stretching, the loss of information is recovered by energy correction.

Rest of the paper is organized as follows: Sect. 2 discussed a state-of-the-art for the problem, Sect. 3 proposed a method for contrast enhancement and energy correction, results are reported in Sects. 4 and 5 has given concluding remark.

2 Literature Review

Among the literature found for underwater image enhancement, in the very initial stage of research, Bazeille et al. [4] proposed an algorithm which is composed of successive independent steps to enhance underwater images. The method removes moiré effect by frequency domain filtering, corrects nonuniform illumination by Homomorphic filtering, removes Gaussian noise by wavelet denoising, Anisotropic filtering to smooth image by preserving edges, improves contrast by adjusting image intensity values and equalize means of RGB channels. He et al. [5] suggested a method for image dehazing based on a prior called dark channel prior (DCP). In DCP transmission, map is estimated using three color channels red, green and blue. Bianco [6] modified DCP for underwater image. In this method, difference in maximum intensity of red color and maximum intensity between blue and green color is used to estimate the transmission. Drews-Jr [7] proposed another variation in DCP called underwater DCP (UDCP) where transmission map for underwater images is estimated using only blue and green channels. Hazy image formation model of Chiang and Chen [8] use dark channel prior and find scene depth map which is further used to compensate for light scattering and color changes. Aysun Tasyapı Çelebi and Sarp Ertürk [9] proposed enhancement technique for underwater images based on Empirical Mode Decomposition (EMD). In this method enhanced image is reconstructed by combining IMFs using weights estimated by genetic algorithm. Genetic algorithm optimizes entropy and average gradient of image. Iqbal et al. [10] proposed an approach called integrated color model (ICM), which first stretch the contrast of red, green, and blue color channels to equalize the color contrast, then saturation and intensity stretching is performed to correct the color of image and solve lighting problem. Iqbal et al. [11] proposed an approach called unsupervised color correction model (UCM) where von Kries hypothesis is applied to equalize RGB colors, then contrast, saturation and intensity stretching is performed. The method of contrast stretching is further extended by Ghani and Isa [12]. In this method, RGB channels are dual stretched with respect to Rayleigh distribution and then saturation and value components of HSV color model are stretched. Ghani and Isa [13] proposed one more method of dual stretching of RGB channels and are followed by modified von Kries hypothesis for color correction. Sankpal and Deshpande [14] enhanced underwater images by nonuniform illumination correction by local contrast stretching with respect to Rayleigh distribution. Ghani and Isa [15] proposed a method called homomorphic filtering and image fusion (HFIF) to reduce inhomogeneous illumination by using homomorphic filter. Histogram matching is used to balance the color percentage in the image. Two enhanced images are then fused by using wavelet fusion. Then a contrast limited histogram specification is applied to improve contrast. Fang et al. [16] proposed an underwater image enhancement method by fusion. In this method the image is processed by white balance algorithm and algorithm for global contrast enhancement. The enhanced result is computed by weighted fusion, obtained by combining three different weight maps. Ancuti et al. [17] proposed a method where image enhancement takes place by multiscale fusion. In this method, there are three images, the first image is single image dehazed by method by He et al. [5] with patch size 20, the second image is single image dehazed by method by He et al. [5] with patch size 60, and the third image is Laplacian of original image which preserve finest details. Fusion takes place by three weight maps. Ancuti et al. [18] proposed a method of multiscale fusion and consists of two images, the first is gamma-corrected input image and second is sharpened input image. Here, input image is first white balanced.

In the literature above discussed, some methods are computationally complex like in DCP and its variations or methods using fusion technique. Other methods like contrast stretching are computationally simple and can be used for real-time processing, but these methods are using fixed parameters for stretching which may degrades the performance. So, it is necessary to develop a method which is computationally simple and also adaptively selects the parameters.

3 Proposed Method

Methodology in this paper is divided into two parts, the first part is contrast enhancement and the second part is energy correction. The contrast degradation of underwater image is caused by backscattering. Scattering of light in water is wavelength dependant, so the correction method is applied to each color channel separately.

3.1 Contrast Enhancement

By considering all the components of reflected light that appears in the camera plane from Eq. (1) can be written as Eq. (3)

$$I_t = I_d + I_{bs} + I_{fs} \tag{3}$$

In the equation, I_t is total intensity that reached the camera, I_d is direct component without scattering, I_{bs} is backscattered component which does not contribute to the image, and I_{fs} is forward scattered component. After rearranging the components of Eq. (3), we get

$$I_t - I_{bs} = I_d + I_{fs} \tag{4}$$

In Eq. (4), right-hand side ($I_d + I_{fs}$) is the image formed after forward scattering and left-hand side ($I_t - I_{bs}$) is captured image after removing backscattered component from it. As backscattered component does not contribute to the image, we can consider it as a noise. So, right-hand side ($I_d + I_{fs}$) in Eq. (4) gives image after removing noise. But it can be considered as total signal received after scattering (I_{ts}).

$$I_{ts} = I_d + I_{fs} \tag{5}$$

As per Siddiqui [19], the distribution of either amplitude or power of electromagnetic signal received through scattering medium is Rayleigh distributed.

The random variable x is said to be Rayleigh distributed [20] with parameter α if its probability density function is given as

$$fx(x) = \begin{cases} \frac{x}{\alpha^2} e^{-\frac{x^2}{2\alpha^2}} & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

So, the method in this paper apply histogram stretching with respect to Rayleigh distribution to each color channel given by Pratt [21] which was also implemented in [14] as the following:

$$I(c)_{out} = I(c)_{min} + [2 * \alpha(c)^2 * \ln\left(\frac{1}{(1 - p_i(I(c)))}\right)]^{\frac{1}{2}} \tag{7}$$

Here as Eq. (7) is computed over three color channels(c), where c ϵ {R, G, B}, $I(c)_{out}$ is pixel value in transformed image, $I(c)_{min}$ is minimum pixel value in the transformed image, α(c) is parameter value and $p_i(I(c))$ is cumulative distribution function of pixel values of input image.

Adaptive Selection of a Scale Parameter

The Rayleigh distribution as specified by Eq. (6) is completely specified if the parameter α is known, and maximum likelihood estimate is the most efficient estimate of α [19].

The principle of maximum likelihood estimation [22] says that observations x_1, x_2, …, x_n are given, $f_x(x_1, x_2, \ldots, x_n; \theta)$, is a function of θ alone, and the value of θ (or α^2) that maximizes the above probability density function is the most likely value for θ, and it is chosen as its maximum likelihood estimation $\hat{\theta}_{ML}(x)$.

The parameter value in Eq. (7) is estimated for each color channel (R, G, and B) using maximum likelihood estimation as given in [22] using Eqs. (8–16).

Initially, the log-likelihood function is given by

$$l(\theta) \ln \prod_{i=1}^{n} l(x_i; \theta) \tag{8}$$

Here, $I(x_i; \theta)$ is a probability density function of x_i with scale parameter θ.

$$l(\theta) = \ln \prod_{i=1}^{n} (\frac{x_i}{\theta} e^{-x_i^2/2\theta}) \tag{9}$$

$$l(\theta) = \sum_{i=1}^{n} \ln(\frac{x_i}{\theta} e^{-x_i^2/2\theta}) \tag{10}$$

$$l(\theta) = \sum_{i=1}^{n} \ln\, x_i - n \ln\theta - \sum_{i=1}^{n} \frac{x_i^2}{2\theta} \tag{11}$$

Then, the value of θ is calculated such that the log-likelihood function has its maximum value. It can be determined by taking a derivative of log likelihood function with respect to θ and equate it to zero.

$$\frac{\partial}{\partial\theta} l(\theta) = 0 \tag{12}$$

Solving Eq. (12) for θ.

$$\frac{\partial}{\partial\theta}\left\{\sum_{i=1}^{n} \ln x_i - n \ \ln\theta - \sum_{i=1}^{n}\frac{x_i^2}{2\theta}\right\} = 0 \tag{13}$$

$$-\frac{n}{\theta} - \sum_{i=1}^{n}\frac{x_i^2}{2}\left(-\frac{1}{\theta^2}\right) = 0 \tag{14}$$

$$\sum_{i=1}^{n}\frac{x_i^2}{2\theta^2} = \frac{n}{\theta} \tag{15}$$

$$\theta_{ML} = \frac{1}{2n}\sum_{i=1}^{n} x_i^2 \tag{16}$$

θ_{ML} is obtained for all three color channels (R, G, and B), which is estimated value of θ for three channels.

Using this θ_{ML} from Eq. (16), histogram stretching with respect to Rayleigh distribution is performed over input image using Eq. (7). This histogram stretching is applied to each color channel of input image. The resultant image is contrast-enhanced image after removing backscattered component from it.

Effect of Scale Parameter in Rayleigh Stretching

Though some of the other literature [12, 13] uses Rayleigh stretching, selection of scale parameter $\theta(c) = \alpha(c)^2$ of Eq. (7) affects overall contrast of image.

Effect of scale parameter on Rayleigh stretching is shown in Fig. 2. In the Fig. 2, input image is stretched with respect to Rayleigh distribution with scale parameter is equal to 0.4. It is the default parameter value in MATAB function for Rayleigh stretching as well as this value of scale parameter considered by literature [12, 13]. It is seen from the stretched image that some portion (with red circle) is over saturated. Whereas, better results are obtained with scale parameter estimated by our method of estimation. So, mere Rayleigh stretching will not solve the problem of contrast enhancement but selection of scale parameter is equally important. So, it is necessary to find scale parameter manually which is suited for a given image. The other (automatic) method to find appropriate scale parameter is proposed in this paper, i.e., maximum likelihood estimation as it is the most efficient estimate of θ(c) [19].

Fig. 2 Effect of scale parameter θ(c) on image contrast

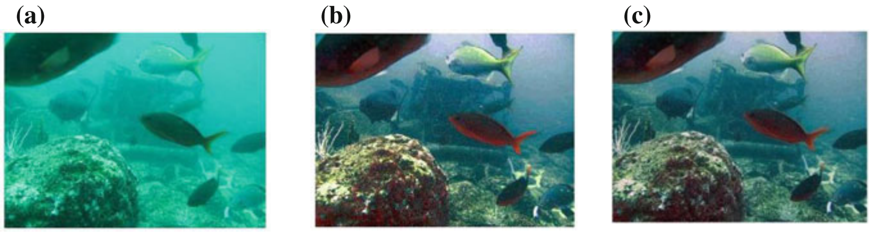


Fig. 3 **a** Original Image (AIE = 6.58) **b** Image after applying contrast enhancement (AIE = 4.46) **c** Image after applying energy correction (AIE = 6.37)

3.2 Energy Correction

In all the literature discussed in Sect. 2, backscattered component is considered as noise as it does not contribute to the image and removed completely. But if we think in other way, this backscattered component may contribute to the image if it does not scatter in backward direction. So, the removal of backscattered component decreases energy content of the image. Removal of backscattered component enhances contrast of the image and contrast-enhanced image generally poor in information content. Energy in the image is given by its mean square value as

$$E_i = \sum_i i^2 p(i) \tag{17}$$

where i is intensity value in image and p(i) is probability of intensity i. Loss of energy is given as

$$E_{loss} = E_{io} - E_{ice} \tag{18}$$

where energy of original image E_{io} and energy of contrast-enhanced image E_{ice} and E_{loss} gives loss. Mean square value for Rayleigh distribution is given as

$$MSV_{Rayeigh} = 2\alpha^2 \tag{19}$$

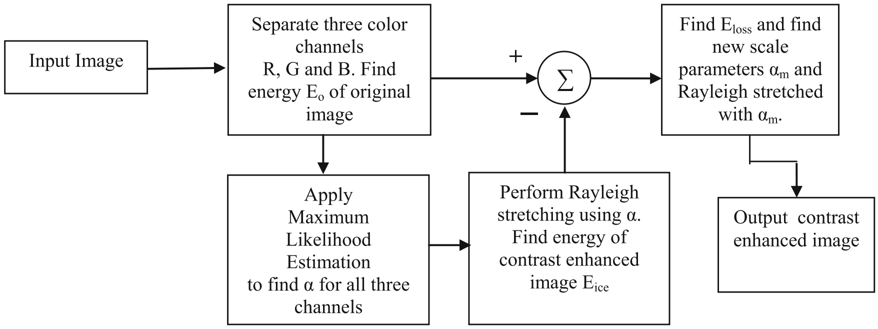


Fig. 4 Process of enhancement

This loss of energy is then recovered by adjusting scale parameter value for each color channel by

$$\alpha_c = \sqrt{\frac{E_{loss}}{2 \times n}} \tag{20}$$

where n gives number of color channels.

The scale parameter for each color channel is then modified as

$$\alpha_m = \alpha_o \times \alpha_c \tag{21}$$

where α_m and α_o are modified and original scale parameter, respectively.

With energy correction, information in image is recovered. Entropy is measure of richness of information in image, so higher value of (AIE) gives image which is rich in information.

3.3 Process of Enhancement

In the process of image enhancement, three color channels of input degraded image are separated. Scale parameter for each color channel is estimated using MLE with Eq. (16), then with estimated scale parameter, each channel is Rayleigh stretched using Eq. (7). Loss of energy is calculated using Eq. (18), and then scale parameter is modified using Eq. (21). The image is again Rayleigh stretched using modified scale parameter. Rayleigh stretching enhances contrast of the image and energy correction improves information contents of image (Fig. 3). The block diagram of complete process is shown in Fig. 4.

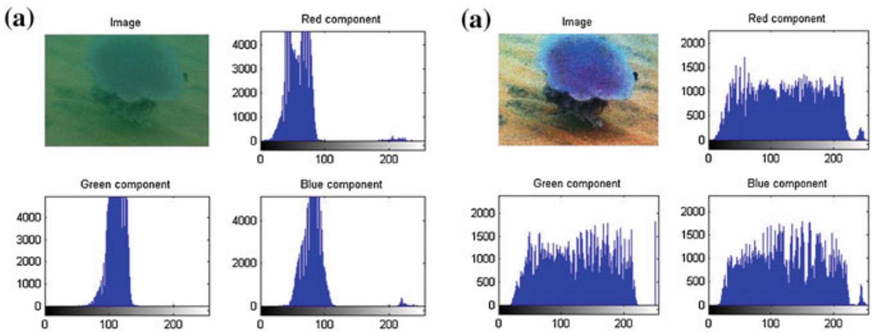
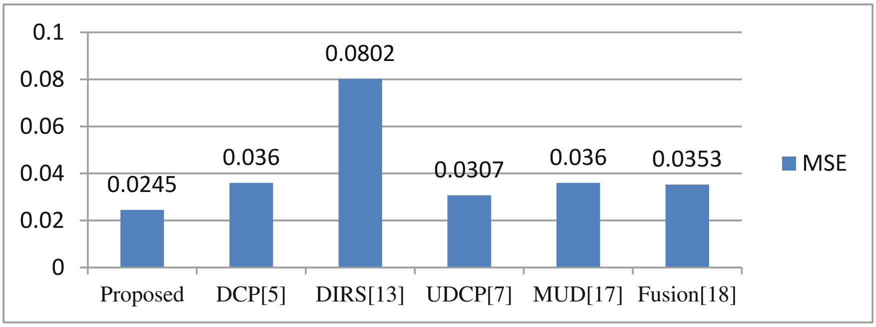


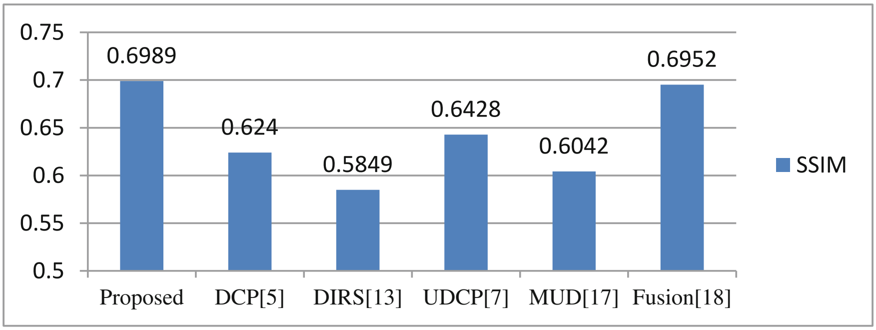
Fig. 5 **a** Input image and histogram of image for red, green, and blue channels, **b** Final image and histogram of image for red, green, and blue channels

4 Results and Discussion

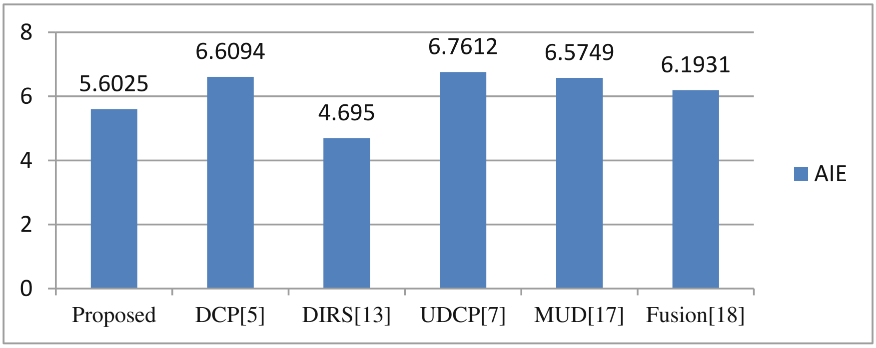
The method proposed in Sect. 3 is divided into two parts. The first part is Rayleigh stretching which enhances the contrast of the image. But the contrast enhancement generally reduces information content of the image. So, the second step is to find loss of information and improve the entropy of image by energy correction of output image. In the Rayleigh stretching, scale parameter which spans complete distribution is adaptively estimated from input image. Figure 5 has given histograms of input image in (a) and final image in (b). It is seen that histograms of input image are concentrated in small portion of intensity levels, i.e., the image is low contrast image. But for the processed image, the histograms are spread over almost all intensity levels, i.e., high contrast image. By just observing histogram of image, it is difficult to compare the performances of the methods. To compare the results of proposed method with state-of-the-art methods, turbid dataset created by Duarte et al. [23] is used in this paper. This dataset is created to give images with the effect of haze which is generally created by backscattering. As the ground truth is available with the dataset, it is easy to compare the performance of the methods. The methods are compared using mean square error (MSE) [12], Structural SIMilarity index (SSIM) [24], and Average Information Entropy (AIE) [12]. Average values for three parameters for 19 images in the dataset are shown in graphs given in Fig. 6.



(a)



(b)



(c)

Fig. 6 Quantitative comparison of results **a** mean square error, **b** structural similarity index, **c** average entropy

5 Conclusion

The algorithm devised in this paper, has given solution to the problem of backward scattering and information loss in underwater images. The results of proposed methods are compared with results of state-of-the-art methods using three parameters. From the comparison, it is seen that proposed method has given best results for two parameters MSE and SSIM, whereas AIE value is smaller for some methods in literature. So, the performance of proposed method is better to recover the original image and the structural components in it. Whereas, entropy of images are still less than some other methods. The improvement in contrast and SSIM index of the image helps to clearly identify the objects in images which make further processing of underwater images easier. The proposed method gives overall best results for given data set.

References

1. Jaffe, J.S.: Computer modelling and the design of optimal underwater imaging systems. IEEE J. Ocean. Eng. **15**(2), 101–111 (1990)
2. Funk, C., Bryant, S., Heckman, P.: Handbook of underwater imaging system design. Technical report TP303, Naval Undersea Centre, San Diego, Calif, USA (1972)
3. Schettini, R., Corchs, S.: Underwater image processing: state of the art of restoration an image enhancement methods. EURASIP J. Adv. Signal Process. **2010** (2010)
4. Bazeille, S., Quidu, I., Jaulin, L., Malkasse, J.P.: Automatic underwater image pre-preprocessing. In: Proceedings of the Caracterisation du Milieu Marin (CMM '06), Brest, France, Oct 2006
5. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. In: IEEE Computer society Conference on Computer Vision and Pattern Recognition, pp. 1319–1321 (2009)
6. Carlevaris-Bianco, A., Eustice, M.R.: Initial results in underwater single image dehazing. OCEANS 2010, pp. 1–8 (2010)
7. Drews, P Jr., Nascimento, E.R., Boteiho, S., Campos, M.: Underwater depth estimation and image restoration based on single images. IEEE Comput. Graph. Appl. **36**(2) (2016)
8. Chiang, J.Y., Chen, Y.: Underwater Image Enhancement by Wavelength Compensation and Dehazing (WCID). IEEE Trans. Image Proc. **21**(4), 1756–1769 (2012)
9. Çelebi, A.T., Ertürk, S.: Visual enhancement of underwater images using empirical mode decomposition. Expert Syst. Appl. **39**, 800–805 (2012)
10. Iqbal, K., Salam, R.A., Osman, A., Talib, A.Z.: Underwater image enhancement using integrated color model. IAENG Int. J. Comput. Sci. **34**, 2 (2007)
11. Iqbal, K., Odetayo, M., James, A., Salam, R.A., Talib, A.Z.: Enhancing the low quality images using unsupervised color correction method. In: International Conference on System Man and Cybernetics (SMC), Istanbul, pp. 1703–1709, 10–13 Oct 2010
12. Abdul Ghani, A.S., Isa, N.A.M.: Underwater image quality enhancement through Rayleigh-stretching and averaging image planes. Int. J. Naval Arch. Ocean Eng. **6**(4), 840–866 (2014)
13. Abdul Ghani, A.S., Isa, N.A.M.: Underwater Image Quality Enhancement Through Composition of Dual Intensity Images and Rayleigh-stretching, vol. 3, p. 757. Springer Plus (2014)
14. Sankpal, S.S., Deshpande, S.S.: Nonuniform illumination correction algorithm for underwater images using maximum likelihood estimation method. J. Eng. Dec 2015 (2016)

15. Ghani, A.S.A., Isa, N.A.M.: Homomorphic filtering with image fusion for enhancement of details and homogeneous contrast of underwater image. Indian J. Geo-Mar. Sci. **44**(12), 1904–1919 (2015)
16. Fang, S., et al.: Effective single underwater image enhancement by fusion. J. Comput. **8**(4), 904–911 (2013)
17. Ancuti, C., Ancuti, C.O., De Vleeschouwer, C., Garcia†, R., Bovik, A.C.: Multi-scale underwater descattering. In: 23rd International Conference on Pattern Recognition (ICPR) 2016, Cancun, Mexico (2016)
18. Ancuti, C.O., Ancuti, C., De Vleeschouwer, C., Bekaert, P.: Color balance and fusion for underwater image enhancement. IEEE Trans. Image Process. **27**(1), 379–393 (2018)
19. Siddiqui, M.M.: Some problems connected with rayleigh distributions. J. Res. Nat. Bur. Stand.-D. Radio Propag. 66D(2), 167–174 (1962)
20. Papoulis, A., UnnikrishnaPillai, S.: The Concept of a Random Variable in Probability, Random Variables and Stochastic Processes, 4th edn. Tata McGraw-Hill Publishing Company Limited, New Delhi, Twelth Print (2007)
21. William, K.P.: Image enhancement. In: Digital Image Processing, 3rd edn. Wiley (2001)
22. Papoulis, A., UnnikrishnaPillai, S.: ‘Statistics’. In: Probability, Random Variables and Stochastic Processes. Tata McGraw-Hill Publishing Company Limited, New Delhi, 4th edn. Twelth Print (2007)
23. Duarte, A., Codevilla, F., De Gaya, J.O., Botelho, S.S.C.: A dataset to evaluate underwater image restoration methods. In: Proceedings of the IEEE Conference on OCEANS 2016, Shanghai, China, Apr 2016
24. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 1–14 (2004)

Medical and Color Image Compression with Fractal Quadtree with Huffman Coding for Different Threshold Values

Sandhya Kadam and Vijay Rathod

Abstract Fractal Image Compression (FIC) is characterized by long encoding time and high Compression Ratio (CR). Further, as medical images being voluminous, a high CR is required to reduce the storage space. Fractal Image compression adopts affine transforms. In view of this, the present paper aims in providing an implementation of a hybrid approach by combining Quadtree fractal with Huffman coding with different threshold values and a comparative analysis of the different types of input images such as color as well as different modalities of medical images as MRI and X-ray to achieve high CR by still retaining the quality of the image. The implementation is carried out and results are obtained using MATLAB. The performance parameters as encoding time, compression ratio PSNR, and decoding time are compared. The results have shown that with an increase in threshold value, CR increases with a decrease in image quality for color as well as medical images.

Keywords Fractal Quad-tree decomposition · Hybrid methodology · Medical image compression · Threshold

1 Introduction

There are different lossless and lossy compression techniques. Joint Photographic Expert Group (JPEG) and Fractal Image Compression (FIC) are examples of lossy compression. Fractal image compression [1] locates self-similar sections of an image, then using a fractal algorithm to generate the sections, performs compression, and fractal codes are transmitted. The fractal codes are applied on any initial image at

S. Kadam (✉)
Faculty of Engineering, Pacific Academy of Higher Education and Research, Udaipur, Rajasthan, India
e-mail: kadam.sandhya@gmail.com

V. Rathod
Head of Department, St. Xavier's Technical Institute, Mumbai, Maharashtra, India
e-mail: vicky7574@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_96

the receiver to get an original image. Fractal code consists of coordinates of range block, affine transformation, and intensity difference.

2 Literature Review

A review on fractal-based image coding methodology is available in the literature [1]. The encoding process consists of approximating the small image blocks, called range blocks, from the larger blocks, called domain blocks of the image. The process of fractal image coding is to find the appropriate domain block for each range block using Iterated function system (IFS) mapping. An improved partitioned Iterated Function System [2] is also introduced to convert RGB color space into YUV color space for an input color image. Fractal-based techniques are studied [3] [4]. Fractal techniques can be applied to medical images [5–7] as medical image compression is a challenging field. The fractal techniques can be enhanced [8]. The structure similarity of fractal can be improved [9].

Recently, fractal techniques are based on quantum theory [9] and variation in block size of fractal methods [10]. The comparative study is carried out for compression of medical images [11, 12]. Quad-tree methods are also studied [13–17]. Hence, Quadtree with Huffman Coding is proposed for medical image compression with varying the threshold value.

3 Proposed System

Quad-Tree Decomposition and Huffman Coding (QDHC) for different threshold values are proposed for medical images. Quad-tree Decomposition is one of the partition-based methods. It divides an image into variable size range block. In this type of partition, a square image is split into square blocks of equal sizes, and then tests each block to check whether each block meets criteria of homogeneity. It is represented in a tree-like structure, where each node will have four subnodes. Adjustment of Quad-tree size is done by using two parameters, minimum level and maximum level. By this method, it is possible to increase the compression ratio and reduce the bits used to represent an image, i.e., bits per pixel (bpp). Huffman coding method was introduced by D.A. Huffman. This is a variable length coding and used to remove the redundancy in the image. In this algorithm, the probability of all alphabet symbols are arranged in decreasing order. The symbols with small frequency will have long codewords and vice versa.

The following steps are given to implement proposed algorithm:

1. Quad-tree partitioning is used to divide the input image using threshold, minimum Dimension and maximum dimension of 2 and 64, respectively.
2. Record x, y, mean value, and block size from Quad-tree Decomposition.

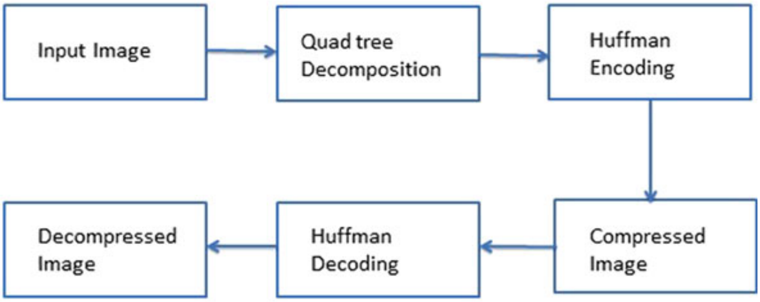


Fig. 1 QDHC fractal compression technique

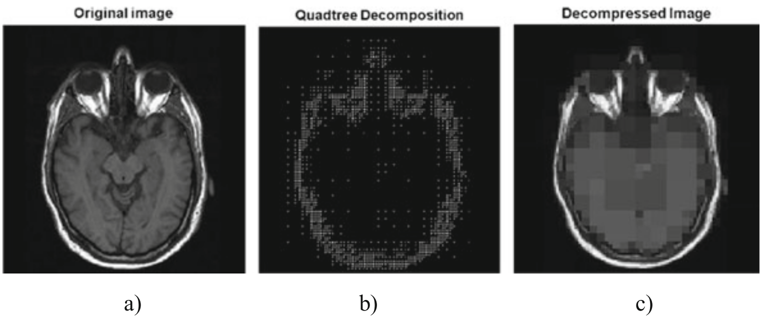


Fig. 2 **a** Original MRI image. **b** Quad-tree decomposition. **c** Decompressed image

3. After Quad-tree decomposition, Huffman coding is applied to complete encoding of the image and then calculate the CR.
4. Once image is reconstructed from fractal codes, apply Huffman decoding to reconstruct the image and calculate PSNR. The block diagram is shown in Fig. 1.

4 Experimental Results

The images are from standard image database as SIPI [18]. The results are given for color as well as medical images with different modalities (Fig. 2).

The parameters as encoding time, compression ratio, PSNR and decoding time are recorded by varying the threshold value as 0.01, 0.1, 0.2, 0.4, and 0.99 for the proposed system. As threshold value increases, encoding time reduces and compression ratio increases with decrease in image quality. Also, with increase in threshold value, decoding time reduces (Tables 1, 2, and 3).

Table 1 The performance parameters as threshold, encoding time(s), CR, PSNR, and decoding time(s) for QDHC with input as color images

Color Images	Threshold	Encoding time(s)	CR	PSNR (dB)	Decoding time(s)
Pepper.tiff	0.01	3.99	1.24	29.59	176.82
	0.1	1.33	2.67	29.09	83.93
	0.2	2.47	4.08	27.88	53.45
	0.4	1.40	9.69	23.96	21.83
	0.99	0.62	969	15.62	0.55
Lena.tiff	0.01	3.74	1.27	30.20	162.76
	0.1	2.16	2.76	29.64	75.64
	0.2	1.74	4.41	28.15	50.08
	0.4	1.32	12.99	23.34	15.09
	0.99	0.50	908.64	16.92	0.54
Mandrill.tiff	0.01	3.64	1.23	26.60	166.02
	0.1	3.07	1.52	26.49	135.41
	0.2	2.39	2.24	25.86	90.15
	0.4	1.17	13.01	22.64	14.46
	0.99	0.47	937.90	19.02	0.50
Sailboat.tiff	0.01	3.67	1.27	27.21	174.24
	0.1	2.54	2.08	27.00	106.79
	0.2	2.19	2.83	26.32	78.17
	0.4	1.68	5.99	23.98	35.11
	0.99	0.52	937.90	15.70	0.49

Table 2 The performance parameters for QDHC with input as MRI images

MRI Images	Threshold	Encoding time(s)	CR	PSNR (dB)	Decoding time(s)
MRI-t1-1.tiff	0.01	3.63	2.00	27.90	145.9
	0.1	2.76	3.41	27.75	75.26
	0.2	2.25	4.79	27.19	53.14
	0.4	1.8	8.82	24.81	26.91
	0.99	1.01	68.20	18.22	2.14
MRI-t1-2.tiff	0.01	3.56	2.12	27.10	120.3
	0.1	2.56	3.40	26.99	72.99
	0.2	4.1	4.69	26.52	52.44
	0.4	1.87	8.17	24.06	28.97
	0.99	1.04	68.54	18.34	2.16
MRI-t2-1.tiff	0.01	3.30	1.76	25.66	158.38
	0.1	2.82	2.68	25.61	105.66
	0.2	2.7	3.25	25.40	77.96

(continued)

Table 2 (continued)

MRI Images	Threshold	Encoding time(s)	CR	PSNR (dB)	Decoding time(s)
	0.4	2.19	4.99	25.16	61.63
	0.99	1.49	65.02	15.75	3.20
MRI-t2-2.tiff	0.01	4.31	1.94	26.59	152.37
	0.1	3.31	2.83	26.51	104.35
	0.2	3.09	3.65	26.13	80.06
	0.4	2.07	5.91	24.60	42.28
	0.99	0.85	169	16.17	0.86
MR_-t2-3.tiff	0.01	3.24	2.20	27.04	118.6
	0.1	2.6	3.14	26.97	80.08
	0.2	2.36	3.97	26.64	63.19
	0.4	2.01	6.33	25.02	38.89
	0.99	0.92	192.61	15.77	0.74

Table 3 The performance parameters for QDHC with input as X-Ray images

X-Ray Images	Threshold	Encoding time(s)	CR	PSNR (dB)	Decoding time(s)
Foot.tiff	0.01	3.59	2.77	26.68	102.87
	0.1	2.59	5.38	26.57	48.31
	0.2	1.95	7.49	26.25	33.29
	0.4	1.73	10.45	25.23	23.43
	0.99	0.91	152.94	15.12	1.04
Chest.tiff	0.01	3.94	1.54	32.04	172.92
	0.1	2.86	2.83	31.15	93.55
	0.2	1.97	6.68	28.55	38.44
	0.4	1.53	20.03	24.08	11.5
	0.99	0.76	188.66	12.22	0.765
Hand.tiff	0.01	2.34	3.46	32.68	69.92
	0.1	1.94	6.01	32.41	39.44
	0.2	1.66	9.41	30.50	24.69
	0.4	1.44	27.33	25.04	7.65
	0.99	0.73	258.77	17.78	0.67
Hand1.tiff	0.01	3.24	1.91	32.54	134.28
	0.1	2.25	3.99	31.94	63.46
	0.2	1.74	8.82	27.55	27.21
	0.4	1.28	32.72	22.37	6.06
	0.99	0.72	250.85	17.75	0.66

Table 4 Comparative analysis of existing transform-based methods with the proposed method with threshold of 0.2 for MRI image

Methods	Encoding time(s)	CR	PSNR (dB)	Decoding time(s)
Proposed method	2.25	4.79	27.19	53.14
DCT	1.57	28	28.5	0.9
DWT(Haar)	8	45	62	0.8
DWT(EZW)	17	2	71	91
FIC	6.9	5.8	33	0.59

5 Comparative Analysis of Performance of Proposed Method

Medical image compression is challenging. Here, the compression is achieved for medical images as MRI and X-Ray. The proposed method has less encoding time as compared to other methods. But DCT method has the lowest value of encoding time even than the proposed system. The quality of reconstructed image of other than proposed system is superior. Even decoding time of proposed method is larger than other methods. The compression achieved with proposed method is lesser than fractal method but higher than DWT with EZW (Table 4).

6 Conclusion

Fractal Quad-tree coding with Huffman coding with different values of threshold can be applied to color as well as medical images. Here, fractal Quadtree with Huffman coding is implemented successfully for color as well as different modalities of medical images as MRI and X-Ray. With this proposed method, compression is also achieved for medical images. The threshold value of Quad-tree decomposition is varied and compression is achieved for color as well as medical images as MRI and X-ray. As threshold value increases, compression ratio of the proposed method increases with decrease in image quality.

In future, structure similarity of fractal can be exploited using better techniques. The different fractal techniques can be applied for overlapping range blocks.

References

1. Jacquin, A.E.: Fractal image coding: a review: In: Proceedings of the IEEE, vol. 81, pp. 1451–1465 (1993)
2. Eman, A., Loay, E.: George: study of fractal color image compression using YUV component, In: Proceedings of the IEEE 36th International Conference on Computer Software and Applications (2012)
3. Sandhya, D.K., Vijay, R.R.: Fractal based image compression techniques: Int. J. Comput. Appl. **178**(1), (2017)
4. Sandhya, K., Vijay, R.R.: DCT with quad tree and Huffman coding for color images. Int. J. Comput. Appl. **173**(9), (2017)
5. Sandhya, K., Vijay, R.R.: Fractal Coding for Texture, Satellite and Grayscale images to reduce Searching time and Complexity. In: Intelligent Engineering Informatics, p. 695. Springer (2018)
6. Padmashree, S., Nagpadma, R.: Statistical analysis of objective measures using fractal image compression. In: Proceedings of the IEEE International Conference on Signal and Image Processing Applications, pp. 563–568, (2015)
7. Rupa, S., Mohan, V., Venkatraman, Y.: MRI brain image compression using spatial fuzzy clustering technique. In: Proceedings of the IEEE International Conference on Communication and Signal Processing, India, pp. 915–918 (2014)
8. Mahalaxmi, G.V.: Implementation of image compression using fractal image compresssion and neural network for MRI images: In: Proceedings of the IEEE International Conference on Information Science(ICIS), pp. 60–64 (2016)
9. Al-saidi, N.M.G., Aquil, H.A.: Towards enhancing of fractal image compression via block complexity. In: Proceedings of the IEEE Annual Conference on New Trends in Information and Communication Technology Applications (NTICT), pp. 246–251(2017)
10. Suvidhya, K., Priyanka, A., Rohtash, D.: Improved structure similarity in fractal image compression with quad tree. Indian J. Sci. Res. 46–50 (2017)
11. Padmashree, S., Naga Padma, R.: Different approaches for implementation of fractal image compression on medical images: In: Proceedings of the IEEE International Conference on Electrical, Electronics, Communication and Optimization Techniques(ICEECCOT), pp. 66–72 (2016)
12. Yang, Y., Bai, G., Chiribella, G., Hayashi, M.: Compression for quantum population coding. In: IEEE International Symposium on Information Theory (ISIT), pp. 1973–1977 (2017)
13. Abdul, N., Salih, J.: Fractal coding technique based on different block size. In: IEEE Al. Sadeq International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA), pp. 1–6, (2016)
14. Padmashree, S., Nagapadma, R.: Comparative analysis of JPEG compressionand fractal image compression for medical images. Int. J. Eng. Sci. Technol. **5**(11), 1847–1853 (2013)
15. Bhavani, S., Thanushkodi, K.G.: Comparison of fractal coding methods for medical image compression. In: IET Image Processing, vol. 7(7), pp. 686–693 (2013)
16. Veenadevi, S., Ananth, A.G.: Fractal image compression using quad tree decomposition and Huffman coding. Int. J. Signal Image Process. **3**(2) 207–212 (2012)
17. Pinki, Rajesh Mehra: Quad tree decomposition based image analysis using intensity difference method **10**(7), 1–10 (2016)
18. SIPI image database. https://www.sipi.usc.edu

A Novel Method to Detect Fovea from Color Fundus Images

Samiksha Pachade, Prasanna Porwal and Manesh Kokare

Abstract The computer-aided diagnosis technology in retinal image analysis requires localization of different fundus structures. Efficient detection and localization of fovea are essential in the analysis of diabetic macular edema. This paper demonstrates a novel technique for detection of fovea from color fundus images based on image enhancement by adaptive manifold filter and further mathematical morphological operations for final foveal center localization. The major advantage of the proposed technique is that it does not need a spatial relationship of optic disc and vessels for the detection of fovea. It is robust to illumination changes and interference caused by retinal pathologies. Experiments show encouraging results that are analyzed on five publically available databases DRIVE, HEI-MED, DIARETDB1, HRF, and MESSIDOR with an accuracy of detection as 100%, 99.40%, 98.88%, 100%, and 98.66%, respectively. Comparative analysis of results indicates that the proposed method achieves better performance than other earlier methods present in the literature.

Keywords Macula detection · Fovea detection · Retinal image analysis
Diabetic macular edema

1 Introduction

The retina is a semi-transparent light-sensitive layer present at back of the eye. Figure 1 illustrates the fundamental structures in color fundus image like macula, fovea, optic disc (OD), and blood vessels (BV). The fovea is a dark-shaded region

S. Pachade (✉) · P. Porwal · M. Kokare
Center of Excellence in Signal and Image Processing, Shri Guru Gobind Singhji
Institute of Engineering and Technology, Nanded, India
e-mail: 2017pec601@sggs.ac.in

P. Porwal
e-mail: porwalprasanna@sggs.ac.in

M. Kokare
e-mail: mbkokare@sggs.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_97

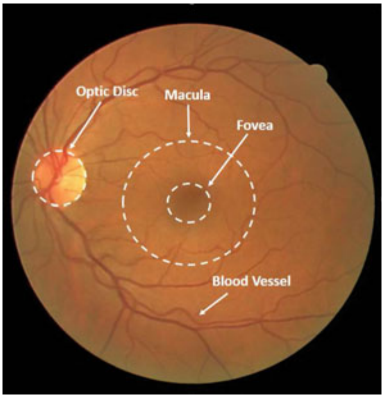


Fig. 1 Fundamental components of fundus image

which is oval in shape having diameter equal to 1.5 mm and it is nearly equal to the diameter of optic disc. Fovea is the most sensitive region with majority of photoreceptors present. It is responsible for central, most accurate and color vision. Any deposit on the foveal region creates blindness to central vision.

Diabetes mellitus (DM) is a problem with greater significance and huge financial burden on the global economy with its incidence is growing at epidemic levels [1]. Diabetes leads to complications such as diabetic retinopathy (DR). DR is the prevalent cause of vision loss in the working-age population. Patients with DR can get affected by diabetic macular edema (DME) at any time during its progression. The leakage of blood vessels causes accumulation of fluid in the macula region is responsible for DME. Severity grading of DME is based on presence of accumulated fluid corresponding to the distance from fovea center. Hence, automatic detection of fovea is essential in the development of computer-aided diagnosis (CAD) system for DR and DME. Computer-aided detection of edema is essential for early diagnosis of disease.

2 Related Work

Identification of fovea has got less consideration; this is likely because fuzzy appearance and low contrast of fovea with its background make its detection difficult and challenging. There are few methodologies presented in the literature.

Tobin et al. [2] proposed technique for fovea localization. Optic disc is located and vascular arch is estimated using parabolic model. Location of fovea is decided using these two anatomical structures. Sagar et al. [3] detect OD using principle component analysis (PCA) and blood vessels are mask out using mathematical morphologic operation. Fovea is determined using 1% of the darkest pixels. Niemeijer et al. [4] utilized point distribution model along with cost function to estimate the location of fovea. Welfer et al. [5] used morphological operation on region of interest (ROI), which is determined using the relationship of OD and fovea. Chin et al. [6] made use of minimum vessel density region for the localization of fovea center. Vessel density is calculated using binary vessel mask. Akram et al. [7] determined macula using the feature set and then classify it as macula or non-macula using Gaussian mixture model classifier. Feature set is determined by detecting all dark regions from retina. Aquino et al. [8] detected fovea center using the positional statistics with respect to OD and vessel tree. Kao et al. [9] locate OD by template matching and then fovea is searched in vessel free region. Medhi and Dandapat [10] utilized intensity information of processed red plane of color fundus image for localization of fovea. Molina-Casado et al. [11] made use of template matching technique along with correlation for the detection of macula blob. Tan et al. [12] used seven-layer convolutional neural network (CNN) to classify each pixel of retinal image as background, blood vessels, OD, and fovea.

We proposed a novel and robust technique for automatic fovea detection, which can be used in automatic DR screening projects. The algorithm initially searches for the fovea region with prior information that it is the dark intensity region. Lastly, it segments the fovea using the morphological operations. The major advantage of the presented methodology is that it does not require prior identification of optic disc, blood vessels, and/or any other fundus pathologies, thus reducing the complexity in identification. The accuracy of the technique is tested on five publicly accessible retinal datasets.

The rest of the paper is organized as follows; Sect. 3 describes proposed method, the result is discussed in Sects. 4 and 5 give the conclusion.

3 Proposed Method

3.1 Field of View Generation

The green plane $g = (g(i))_{i \in \Omega}$ is defined over $\Omega \in \mathbb{R}^2$. Figure 2a, b shows the color fundus and its g plane. Green plane shows better contrast between foveal region and background as compared to red and blue plane. Thus, for the further analysis, green

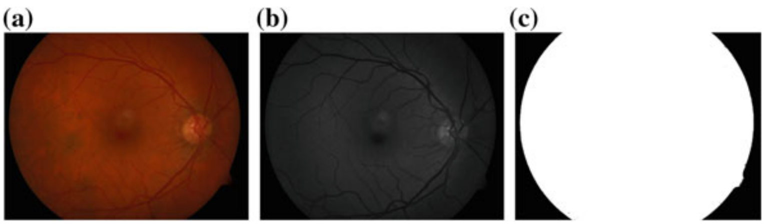


Fig. 2 **a** Color fundus image, **b** green channel, **c** field of view (FOV)

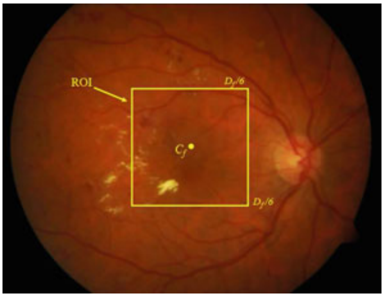


Fig. 3 ROI from retina

plane is considered. The field of view (FOV) as shown in Fig. 2c is determined by using mean estimate of g as follows:

$$FOV = \begin{cases} 1, & g(i) \geq g(mean)/2 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

3.2 ROI Selection

Utilizing the FOV mask, FOV center $C_f = \frac{FOV_{width}}{2}, \frac{FOV_{height}}{2}$ and FOV diameter $D_f = mean([FOV_{width}\ FOV_{height}], 2)$ are derived. The ROI center is same as the C_f and ROI is selected by considering $D_f/6$ part of image from C_f as shown in Fig. 3. To reduce computational complexity, ROI is further downsampled by factor of 2. The output image is given by I_{ROI}.

3.3 Image Enhancement

Uneven illumination in the retinal atlas caused while capturing the fundus image poses certain challenges in the identification of retinal structures including fovea. Hence, ROI enhancement is done in order to create ease in fovea detection. It is done by utilizing adaptive manifold filter [13] for the purpose which has the capability to accelerate filtering and deliver high-quality enhanced image. In this filter, weighted average of the input gray scale pixels is done and a new set of pixel gray scale is produced. The weights of the filter are given by kernel ϕ. Filtering the image with ϕ gives a new image I_{enhc}:

$$I_{enhc_i} = \frac{\sum_{n_j \in s} \phi(\hat{n}_i - \hat{n}_j) I_{ROI_i}}{\sum_{n_j \in s} \phi(\hat{n}_i - \hat{n}_j)} \tag{2}$$

Functional selection of ϕ is an axis-aligned Gaussian, as given below:

$$\phi \sum(\hat{n_i} - \hat{n_j}) = exp(-\frac{1}{2}(\hat{n_i} - \hat{n_j})^T \sum^{-1}(\hat{n_i} - \hat{n_j}) \quad (3)$$

where $\sum$ is a diagonal covariance matrix of size $d \times d$. The $\hat{n_i}$ and $\hat{n_j}$ are neighborhoods around the corresponding pixel.

3.4 Fovea Detection

Subtract the image I_{enhc} from I_{ROI} and multiply with detailed amplification factor A as

$$I_{ampl} = I_{ROI} + (I_{ROI} - I_{enhc}) \times A \quad (4)$$

where the value of A is equal to 10. Image is filtered using multidimensional Gaussian filter where input array values outside the bounds of the array are assigned to the nearest array border. Mapping of intensity values of the image obtained by applying a multidimensional filter to the newly obtained values in such a way that 1% of data is saturated to low and high intensities. The output image is as shown in Fig. 4c. The output image so obtained is denoted by I_{ampl}.

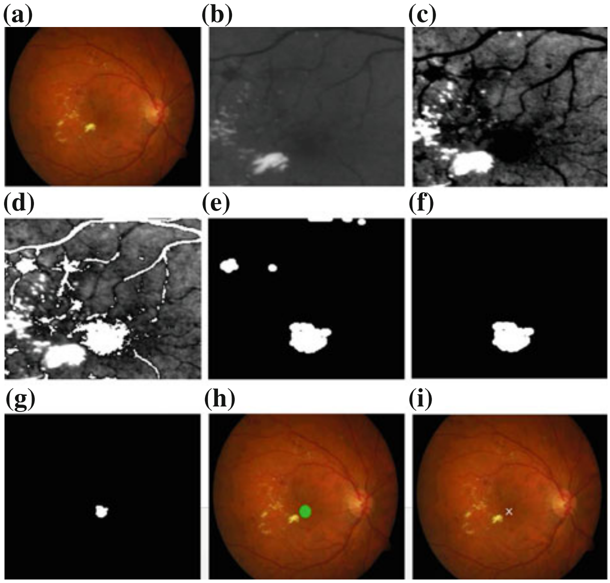


Fig. 4 **a** Original image, Image after: **b** ROI selection, **c** enhancement, **d** thresholding, **e** morphological opening, **f** selecting largest connected component, **g** mapping to original size, **h** fovea detection, **i** fovea localization

Hard thresholding is done with thresholding intensity equal to zero as shown in Fig. 4d. As intensity of the vessels and dark lesions (i.e., hemorrhages) are also zero, they are also segmented out and output I_{thre} is obtained. Application of morphological opening with a disk-shaped structuring element E with fixed radius of size $s = 4$ is as given in Eq. 5. After opening operation, 95% of small vessels are removed.

$$I_{open}^{Es}(I_{thre}) = \delta^{Es}[\varepsilon^{Es}(I_{thre})] \tag{5}$$

where δ and ε are morphological dilation and erosion operation respectively.

Algorithm 1 Fovea Segmentation and Localization

Input: Original RGB image I and **Parameters** A = 10, s = 4, E = Disc structuring element.
Result: Segmented Fovea ($Fovea$) and its Localization (L_{fovea})
$g(i) \leftarrow I(:, :, 2)$
if $g(i) \geq g(mean)/2$ **then**
    $FOV = 1$
else
    $FOV = 0$
end if
$C_f \leftarrow \frac{FOV_{width}}{2}, \frac{FOV_{height}}{2}$
$D_f \leftarrow mean([FOV_{width}\ \ FOV_{height}], 2)$
$I_{ROI} \leftarrow FindROI(I, C_f, D_f/6)$
$\phi \sum(\hat{n_i} - \hat{n_j}) \leftarrow exp(-\frac{1}{2}(\hat{n_i} - \hat{n_j})^T \sum^{-1}(\hat{n_i} - \hat{n_j})$
$I_{enhc} \leftarrow I_{enhc_i} = \frac{\sum_{n_j \in s} \phi(\hat{n_i} - \hat{n_j}) I_{ROI_i}}{\sum_{n_j \in s} \phi(\hat{n_i} - \hat{n_j})}$
$I_{ampl} \leftarrow I_{ROI} + (I_{ROI} - I_{enhc}) \times A$
$[p, q] \leftarrow size(I_{ampl})$
if $I_{ampl}(i, j) == 0$ **then**
    $I_{thre}(i, j) = 255$
else
    $I_{thre}(i, j) = I_{ampl}(i, j)$
end if
$I_{open}^{Es}(I_{thre}) \leftarrow \delta^{Es}[\varepsilon^{Es}(I_{thre})]$
$FoveaMask \leftarrow ConnectedComponent(I_{open}, [\,], 1)$
$Fovea \leftarrow resize(FoveaMask)$
$L_{fovea} \leftarrow \frac{Fovea_{width}}{2}, \frac{Fovea_{height}}{2}$

The fovea is the largest connected region and the unwanted detected regions are much smaller than that. All the connected components of I_{open} are detected, and labels are given. To segment out only the fovea, calculate the area of all the labeled connected components and arrange them in descending order, i.e., connected component with the largest area is first one. Only first component is selected as shown in Fig. 4f. Detected fovea is shown in Fig. 4g. Finally, center is detected using regional property. The total algorithm of proposed method is given in Algorithm 1.

Table 1 Comparison of methods present in literature with the proposed method

Method	Need of OD in advance	Need of BV in advance	Database used	Total number of images in dataset	Number of images used for analysis	Success rate for fovea location (%)
Welfer et al. [5]	Yes	No	DRIVE	40	37	100
			DIARETDB1	89	89	96.62
Chin et al. [6]	Yes	Yes	MESSIDOR	1200	303	72.9
Aquino et al. [8]	Yes	Yes	MESSIDOR	1200	1136	98.24
			DIARETDB1	89	89	94.38
Akram et al. [7]	Yes	Yes	MESSIDOR	1200	1200	97.2
			HEI-MED	169	169	98.22
Kao et al. [9]	Yes	Yes	MESSIDOR	1200	1200	97.8
			DIARETDB0	130	130	92.1
			DIARETDB1	89	89	93.1
Medhi et al. [10]	Yes	Yes	DRIVE	40	39	100
			DIARETDB1	89	87	97.70
			MESSIDOR	1200	1190	97.98
			HRF	45	45	100
Molina-Casado et al. [11]	No	No	MESSIDOR	1200	1200	98.58
			DIARETDB1	89	89	100
			ONHSD	99	87	98.85
Proposed Method	**No**	**No**	**DRIVE**	**40**	**40**	**100**
			HEI-MED	**169**	**169**	**99.40**
			DIARETDB1	**89**	**89**	**98.88**
			HRF	**45**	**45**	**100**
			MESSIDOR	**1200**	**1200**	**98.66**

4 Result

Results of the presented algorithm are analyzed on 1543 images of five publicly accessible databases.

From DRIVE [14] 40/40 images detects fovea accurately, 168/169 from HEI-MED [15], 88/89 from DIARETDB1 [16], 45/45 from HRF [17] and 1184/1200 from MESSIDOR [18] database. The presented algorithm for fovea detection resulted in a success rate of 99.38%. The performance of presented method is compared with various methods present in literature and is tabulated in Table 1. Our experimental results are promising and show improvement over previous methods present.

5 Conclusion

In this paper, we have proposed a novel technique for the detection of fovea which is essential for analysis of diabetic macular edema. Furthermore, it can be utilized with vessel detection method to boost the performance of red lesion, retinitis pigmentosa, and/or scar detection. The proposed technique is efficient since there is no need of detecting optic disc and blood vessels. The proposed method can be used in the development of an automatic diagnosis system. In future, it can be incorporated with the detection of fovea and pathologies in the fovea region from optical coherence tomography images for better analysis of DME.

References

1. Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Sahasrabuddhe, V.: Automatic retinal image analysis for the detection of diabetic retinopathy. Biomed. Signal Image Process. Patient Care 146 (2017)
2. Tobin, K.W., Chaum, E., Govindasamy, V.P., Karnowski, T.P.: Detection of anatomic structures in human retinal imagery. IEEE Trans. Med. Imaging **26**(12), 1729–1739 (2007)
3. Sagar, A.V., Balasubramanian, S., Chandrasekaran, V.: Automatic detection of anatomical structures in digital fundus retinal images. In: MVA, pp. 483–486 (2007)
4. Niemeijer, M., Abramoff, M.D., Van Ginneken, B.: Segmentation of the optic disc, macula and vascular arch in fundus photographs. IEEE Trans. Med. Imaging **26**(1), 116–127 (2007)
5. Welfer, D., Scharcanski, J., Marinho, D.R.: Fovea center detection based on the retina anatomy and mathematical morphology. Comput. Methods Program. Biomed. **104**(3), 397–409 (2011)
6. Chin, K.S., Trucco, E., Tan, L., Wilson, P.J.: Automatic fovea location in retinal images using anatomical priors and vessel density. Pattern Recogn. Lett. **34**(10), 1152–1158 (2013)
7. Akram, M.U., Tariq, A., Khan, S.A., Javed, M.Y.: Automated detection of exudates and macula for grading of diabetic macular edema. Comput. Methods Program. Biomed. **114**(2), 141–152 (2014)
8. Aquino, A.: Establishing the macular grading grid by means of fovea centre detection using anatomical-based and visual-based features. Comput. Biol. Med. **55**, 61–73 (2014)
9. Kao, E.F., Lin, P.C., Chou, M.C., Jaw, T.S., Liu, G.C.: Automated detection of fovea in fundus images based on vessel-free zone and adaptive gaussian template. Comput. Methods Program. Biomed. **117**(2), 92–103 (2014)
10. Medhi, J.P., Dandapat, S.: An effective fovea detection and automatic assessment of diabetic maculopathy in color fundus images. Comput. Biol. Med. **74**, 30–44 (2016)
11. Molina-Casado, J.M., Carmona, E.J., García-Feijoó, J.: Fast detection of the main anatomical structures in digital retinal images based on intra-and inter-structure relational knowledge. Comput. Methods Program. Biomed. (2017)
12. Tan, J.H., Acharya, U.R., Bhandary, S.V., Chua, K.C., Sivaprasad, S.: Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network. J. Comput. Sci. **20**, 70–79 (2017)
13. Gastal, E.S., Oliveira, M.M.: Adaptive manifolds for real-time high-dimensional filtering. ACM Trans. Graph. (TOG) **31**(4), 33 (2012)
14. Staal, J., Abramoff, M., Niemeijer, M., Viergever, M., van Ginneken, B.: Ridge based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging **23**(4), 501–509 (2004)
15. Giancardo, L., Meriaudeau, F., Karnowski, T.P., Li, Y., Garg, S., Tobin, K.W., Chaum, E.: Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. Med. Image Anal. **16**(1), 216–226 (2012)

16. Kauppi, T., Kamarainen, J.K., Lensu, L., Kalesnykiene, V., Sorri, I., Uusitalo, H., Kälviäinen, H.: A framework for constructing benchmark databases and protocols for retinopathy in medical image analysis. In: International Conference on Intelligent Science and Intelligent Data Engineering, pp. 832–843. Springer (2012)
17. Khler, T., Budai, A., Kraus, M.F., Odstrilik, J., Michelson, G., Hornegger, J.: Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, pp. 95–100 (2013)
18. Decencière, E., Zhang, X., Cazuguel, G., Laÿ, B., Cochener, B., Trone, C., Gain, P., Ordóñez-Varela, J.R., Massin, P., Erginay, A., et al.: Feedback on a publicly distributed image database: the messidor database. Image Anal. Stereol. **33**(3), 231–234 (2014)

Detection of Malaria Parasite Based on Thick and Thin Blood Smear Images Using Local Binary Pattern

Satishkumar L. Varma and Satishkumar S. Chavan

Abstract Malaria is one of the dangerous diseases transmitted by a female Anopheles mosquito through parasites. Parasite is a type of microorganism. Microscopic examination of blood samples helps to diagnose malaria automatically and faster. It also reduces the time and human errors. This paper aims to experiment and analyze quickly the accurate number of malaria parasites using image processing techniques. Local binary pattern (LBP) technique is used to classify blood smear into thin and thick blood smears. Morphological operations and k-means clustering techniques along with intensity profiles within the cells are used to count infected cells. The experiments are performed over standard datasets using segmentation and morphological operations for thick and thin blood smear images. The performance of the proposed algorithm is evaluated using confusion matrix. The results are compared using sensitivity and specificity. This method proves to be much effective in terms of time considering large rural areas in India.

Keywords Red blood cells · Blood smear · Segmentation · Morphological operation · Malaria Parasite · k-means clustering · Local Binary Pattern

1 Introduction

Infections and spread of diseases due to mosquitoes are real challenges in rural as well as urban areas of the world. Malaria is the most common and dangerous disease caused by a female Anopheles mosquito with the help of parasites. It is a very infectious disease of humans and other animals. It remains one of the most widespread infectious diseases of mankind with 216 million cases worldwide in 91

S. L. Varma
Pillai College of Engineering, Panvel, Navi Mumbai, India
e-mail: varmasl@yahoo.co.in

S. S. Chavan (✉)
Don Bosco Institute of Technology, Kurla (W), Mumbai, India
e-mail: satyachavan@yahoo.co.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_98

countries in 2016, according to the World Health Organization (WHO) [1]. It causes symptoms like fever, fatigue, vomiting, headaches, etc. In few cases, it can cause yellow skin, coma, or death. The symptoms usually begin 10–15 days after being bitten by mosquito. Improper treatment leads to recurrences of the disease after few months which lead to complications in treatment.

The microscopic examination of blood with blood films is used to diagnose malaria. Current state-of-the-art for medical diagnosis and research purposes involves drawing a blood sample from a patient or research subject. This blood sample is smeared onto a slide. It stained in order to color cell nuclei. The mature red blood cell (RBC) does not possess nuclei. This helps the stain to find and marks malarial parasites. These slides are then examined under a microscope. The counting of the number of infected red blood cells is carried out during this examination.

1.1 Related Work

The proposed work is based on semi-automatic diagnosis methods using image processing techniques. This section deals with attempts made by researchers to diagnose malarial parasites using digital image processing (DIP) algorithms. These methods are complex, supervised and need manual intervention or calibration.

The implementation of KNN classifier for diagnosis and screening of malaria is given by Tek et al. [2]. It also suggests color-based features for detecting parasites and non-parasites. The shape-based features are also experimented to analyze the performance improvement. The color, size, and textural features are used to classify parasite infected areas with SVM [3]. Savkare et al. [4] used Otsu thresholding and watershed transform for segmentation followed by SVM classifier using color and statistical features to detect parasites.

Mandal et al. [5] discusses normalized cut algorithm for segmentation using various color spaces. RGB, HSI, and C-Y color spaces are experimented in [6] to detect malaria parasites. This discusses the mathematical model for classification of RBC as infected parasite and non-parasite. The method for counting RBC is also discussed. Comparative analysis of various classifiers for malaria detection along with stage identification is presented in [7]. Extensive literature survey is given in context with features and performance analysis for microscopic images to detect malarial parasites using thin blood smear (TBS) [8].

Tsai et al. [9] proposed technique of automated malaria parasite and infected erythrocytes segmentation (MPIE) which segment microscopic images to detect malaria parasites and infected erythrocytes. The automatic counting of number of malaria parasites using standard image processing algorithms like histogram equalization, thresholding, morphological operations, and connected components analysis is used in [10].

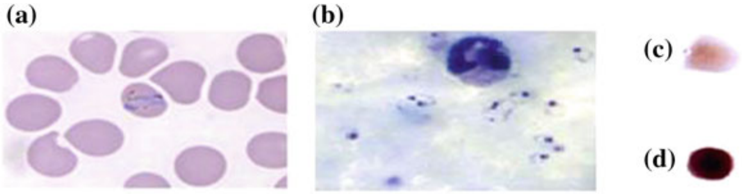


Fig. 1 Blood sample: **a** Infected red blood cell (RBC), **b** Infected RBCs and WBC, **c** Thin blood smear (TBS), **d** Concentrated thick blood smear (CBS)

The proposed research work is presented as: Sect. 2 explains about malaria parasite and diagnostic procedure. Section 3 details the proposed methodology. The experimental results and comparative analysis are given Sect. 4. The conclusion along with future directions is presented in Sect. 5.

2 Malaria Parasites

2.1 Red Blood Cells

The Red Blood Cells (RBC) with malaria infection is declared, if at least one parasite is detected within its interior. White blood cells and free-floating parasites are not considered. The current practice involves manual counting by an expert laboratory technician who can distinguish staining artifacts from actual nuclei, white blood cells, life cycle (depending on specific requirements), and species of malarial parasites. Although manual counting is relatively inexpensive to implement, adequate sensitivity requires proper training and supervision of technicians. This poses problems for both medical care providers in impoverished regions of the world as well as laboratory settings which may benefit from automation of a tedious and time-consuming task.

Automation of such task facilitates laboratory efficiency and also provides an alternative diagnostic tool. Also, expert pathologists are needed in parasite detection and their availability in rural areas is also a challenge. Obviously, automatic malaria parasite detection system is advantageous. The infection in RBC by malaria parasite and its effects are discussed in [11]. Figure 1a shows red blood cell with one infected cell and Fig. 1b shows infected RBCs and single WBC.

Two types of blood smears are preferred in malaria parasite detection and analysis as thin blood smear (TBS) and the concentrated thick blood smear (CBS). TBS drop spreads across a large area of the slide as shown in Fig. 1c. Within 10 min, thin blood smear (TBS) types of images get dried. So, the thin smear needs to be fixed in methanol once these images are dried. It helps to discover the type of species of malaria causing the infection. CBS has certain thickness as it is a drop of blood on

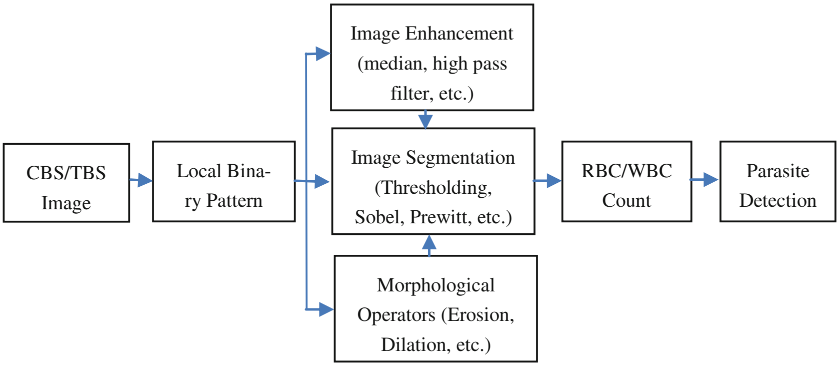


Fig. 2 The proposed methodology for malaria parasite detection

a glass slide as shown in Fig. 1d. These images are normally dried for 30 min. It is also not fixed with methanol. CBS images are useful in detecting infection along with the estimation of parasite concentration.

2.2 Diagnosis Procedure

The diagnosis processes can be manual or automatic. The manual diagnosis is done using microscope which is sensitive and specific. Human intervention is necessary and it takes more time as well as leads to erroneous diagnosis sometimes. Though the microscopy method has an advantage over automatic techniques, some of the problems of manual microscopy can be overcome by automatic process. Image processing techniques are effective in diagnosis of automatic detection of malaria parasites using CBS and TBS images. These automatic diagnosis procedures are unsupervised and highly sensitive. It helps to reduce the false negative cases. It is fast, accurate, and consistent in finding the true malaria cases.

3 Methodology to Detect Malaria Parasite

The proposed technique uses microscopic blood cell images (CBS/TBS) as an input. The source images are classified as CBS or TBS using Local Binary Patterns (LBPs) as a feature descriptor [12]. Various image enhancement techniques are used to improve the quality of image and to reduce noise. The image segmentation is applied followed by morphological operations. The next step is to count number of parasite identified within the image and RBC count of the respective image block as shown in Fig. 2.

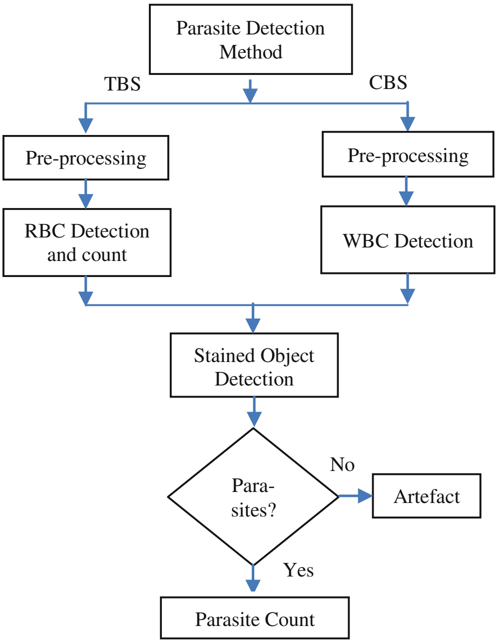


Fig. 3 Malaria parasite count process

Fig. 4 Operation on CBS image: **a** CBS image with malaria parasites and WBC, **b** Otsu thresholding to detect WBC nucleus, **c** Malaria parasites as white spots

The flowchart to enumerate malaria parasite using thick and thin blood smear images is as shown in Fig. 3. After images are classified as thick and thin smears, the counting of RBC/WBCs is required which will be useful for detecting stained cells. The total number of stained cells is the parasite count.

Figure 4a shows CBS image with one WBC and number of malaria parasites. The median filter is used for noise reduction. Otsu thresholding is used as a segmentation technique to separate WBC nucleus as a ROI as shown in Fig. 4b. To detect malaria parasites, k-means clustering is performed. Figure 4c shows the number of parasites in given input CBS image.

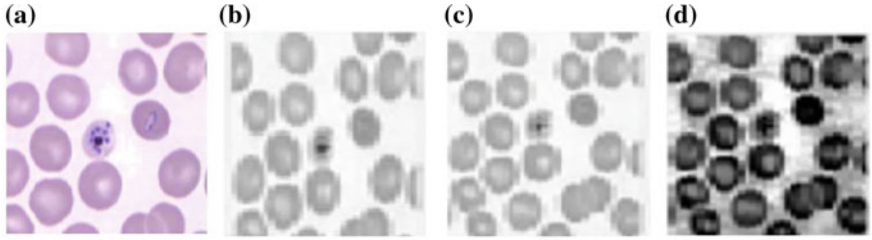


Fig. 5 Operation on TBS image: **a** Thin smear blood (TBS) with two infected cells, **b** RGB to grayscale, **c** filtered image, **d** histogram equalized image

TBS images are well identified, selected, filtered, and reformatted into the desired form. Figure 5a shows TBS image with two infected cells. TBS images sometimes have low brightness, low contrast, and noise. The preprocessing operation plays a significant role in operation on TBS images. The first image is converted into grayscale as shown in Fig. 5b followed by filtering (refer Fig. 5c). Histogram equalization is also preferred preprocessing to improve the contrast of source TBS image as shown in Fig. 5d.

Next step is highlighting ROI and locating objects within image boundaries. Otsu thresholding is used as segmentation tool. Sobel and Prewitt masks are used for edge detection. It detects high-frequency variations and abrupt changes can be highlighted. The morphological operations are used in shape identification based on the structuring elements. The RBC counting helps in detection of malaria parasites. Morphological operations are used for parasite count. The cell dimension is obtained from the step of RBC count. The contour plot helps to find the dimension and count of malarial parasites.

4 Result and Discussion

The dataset includes limited number of TBS and CBS image [13] for experimentation. The thin smear image contains parasites and is in the ring form. The thick smear consists of WBC and parasites. Few images only contain the malaria parasites. The WBC is also absent in few images.

Sensitivity and specificity are used to measure the performance of proposed algorithm. The values for sensitivity and specificity are calculated using four parameter namely true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Sensitivity and specificity are calculated using Eq. 1 and Eq. 2, respectively.

$$Sensitivity = TP/(TP + FN) \tag{1}$$

$$Specificity = TN/(TN + FP) \tag{2}$$

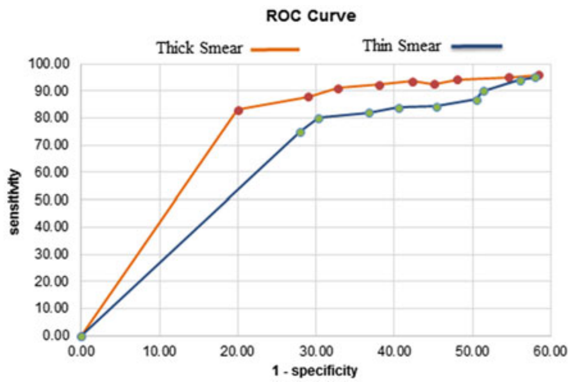


Fig. 6 Performance comparison of malaria parasite detection over thin smear and thick smear

Table 1 Performance for thin blood smear (TBS) images

Metric	Thin smear								Average
	1	2	3	4	5	6	7	8	
TP	85	86	83	81	83	80	82	74	81.750
FP	4	6	6	6	5	5	3	2	4.625
FN	3	5	6	7	6	7	8	15	7.125
TN	3	6	6	7	7	10	7	11	7.125
Accuracy	92.6	88.3	88.1	87.1	88.1	85.3	90.0	87.3	88.362

Table 2 Performance for the concentrated thick blood smear (CBS) images

Metric	Thick smear								Average
	1	2	3	4	5	6	7	8	
TP	60	58	62	62	60	60	61	62	60.625
FP	4	6	7	6	7	8	6	5	6.125
FN	10	12	12	14	13	10	6	6	10.375
TN	3	6	7	8	10	10	10	10	8.000
Accuracy	90.9	85.4	84.1	84.4	81.1	79.5	80.7	81.9	83.515

The ROC curve is plotted using ten samples of thick smear and ten samples of thin smear in Fig. 6. It shows the performance of thick smear (CBS) and thin smear (TBS) images. There is average 91.98% sensitivity for thick smear and 85.51% sensitivity for thin smear.

Table 1 and Table 2 give the values of TP, FP, FN, and TN for eight sets of TBS and CBS images, respectively. It is revealed that average true positive rate (TPR) is 91.98% and 85.51% for thin and thick smears, respectively. However, average false positive rate (FPR) is 40.46% and 44.55% for thin and thick blood smears, respectively. The accuracy of the proposed algorithm for available limited dataset is 88.362% for TBS images and 83.515% for CBS images. Highest value of TPR, FPR, and accuracy are the measures of the better performance of algorithm. Thus, thin smear provides better performance compared to thick smear.

5 Conclusion

Thin blood smear (TBS) and Concentrated thick blood smear (CBS) images are classified using linear binary pattern (LBP). These images are used in the proposed automatic malaria parasite detection system. It helps to identifies number of malaria parasites with the help of image processing algorithms like image segmentation, morphological operations, and edge detection techniques. These techniques together contribute for counting the RBCs, WBCs, and number of infected RBCs by malarial parasites. The performance of the proposed approach for thin blood smear is better as compared to thick blood smear. The system achieves acceptable values of sensitivity and specificity. The automatic detection of malaria parasites is helpful to the patients in rural areas. Further, it is useful to avoid complications due to delayed diagnosis and treatment. The proposed detection of malaria parasites is faster compared to manual process.

The proposed work is experimented with online data sets and using limited cases. It can be tested for its robustness and compared with manual process for large datasets. The work can be extended with other classifiers like Support Vector Machine (SVM), Artificial Neural Network (ANN), Neuro-Fuzzy approaches to analyze, classify, and verify the parasite species based on their size and shapes.

Acknowledgements Authors would like to thank Mr. Parikshit Shembekar, Mr. Niraj Yadav, Mr. Jayajith Jayaprakash, and Mr. Mohammed Shaikh for their help during the implementation of research work.

References

1. World Malaria Report 2017. World Health Organization, Geneva (2017)
2. Tek, F.B., Dempster, A.G., Kale, I.: Computer vision for microscopy diagnosis of malaria. Malar. J. **8**(1), 153 (2009)
3. Linder, N., et al.: A malaria diagnostic tool based on computer vision screening and visualization of plasmodium falciparum candidate areas in digitized blood smears. PLoS ONE **9**(8):e104855, 1–12 (2014)
4. Savkare, S.S., Narote, S.P.: Automatic detection of malaria parasites for estimating parasitemia. Int. J. Comput. Sci. Secur. (IJCSS) **5**(3), 310–315 (2011)
5. Mandal, S., Kumar, A., Chatterjee, J., Manjunatha, M., Ray, A.K.: Segmentation of blood smear images using normalized cuts for detection of malarial parasites. In: 2010 Annual IEEE India Conference (INDICON), Kolkata, India, pp. 1–4 (2010)
6. Abdul-Nasir, A.S., Mashor, M.Y., Mohamed, Z.: Colour image segmentation approach for detection of malaria parasites using various colour models and k-Means clustering. WSEAS Trans. Biol. Biomed. **10**(1), 41–55 (2013)
7. Khot, S.T., Prasad, R.K.: Optimal computer based analysis for detecting malarial parasites. In: Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), pp. 69–80. Springer (2015)
8. Devi, S.S., Sheikh, S.S., Laskar, R.H.: Erythrocyte features for malaria parasite detection in microscopic images of thin blood smear: a review. Int. J. Interact. Multimed. Artif. Intell. **4**(2), 35–39 (2016)

9. Tsai, M.H., Yu, S.S., Chan, Y.K., Jen, C.C.: Blood smear image based malaria parasite and infected-erythrocyte detection and segmentation. J. Med. Syst. **39**(118), 1–14 (2015)
10. Arco, J.E., Gorriz, J.M., Ramirez, J., Alvarez, I., Puntonet, C.G.: Digital image analysis for automatic enumeration of malaria parasites using morphological operations. Expert Syst. Appl. **42**, 3041–3047 (2015)
11. Cowman, A.F., Crabb, B.S.: Invasion of red blood cells by malaria parasites. Cell **124**(4), 755–766 (2006)
12. Varma, S.L., Talbar, S.N.: Video retrieval using local binary pattern. Smart innovation, systems, and technologies. In: Springer Book Series, vol. 31, pp. 123–130 (2015)
13. Sio, S.W.S., et al.: MalariaCount: an image analysis-based program for the accurate determination of parasitemia. J. Microbiol. Methods **68**, 11–18 (2007)

Gender Identification from Frontal Facial Images Using Multiresolution Statistical Descriptors

Prabha, Jitendra Sheetlani, Chitra Dhawale and Rajmohan Pardeshi

Abstract Gender identification is a significant task which is very useful in many computer applications like human–computer interaction, surveillance, demographic studies, and forensic studies. Being one of the most popular soft biometrics, gender information plays a vital role in improvement of the accuracy of biometric systems. In this paper, we have presented an approach based on multiresolution statistical descriptors derived from histogram of Discrete Wavelet Transform. First, the input facial image was enhanced by applying contrast limited adaptive histogram equalization. During feature extraction, multiresolution statistical descriptors were computed and fed into the Nearest Neighbor, Support Vector Machine, and Linear Discriminant Analysis classifiers respectively. We have achieved encouraging accuracy for gender identification on complex dataset of frontal facial images.

Keywords Discrete wavelet transform · Statistical features · Gender identification · Face image · Support vector machine

1 Introduction

The biometrics and multimodal biometrics plays a significant role in authenticating a person. The use of biometrics is unavoidable in present scenario, because, the use of technology has tremendously enhanced and reached almost all the corners of human life. The authenticity of a person/individual is the prime concern of many Governemnt/Private Institutions. Particularly, forensic department needs a robust tool to authenticate a person/individual. However, a common man yet to ensure

Prabha · J. Sheetlani
Department of Computer Science, SSSUTMS, Sehore, Bhopal, India

C. Dhawale
P. R. Pote College of Engineering and Management, Amravati, India
e-mail: dhawalechitra@gmail.com

R. Pardeshi (✉)
Department of Computer Science, Karnatak College, Bidar, India
e-mail: madhurrajmohan1@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_99

that the use of latest technologies for his/her day-to-day transactions are secure. Most of the recent technologies are using biometrics for user authentication. In this context, to make the commercial applications more reliable and fast to work on large databases, which needs a robust database indexing technique. Gender is an important demographical attribute and plays a significant role for indexing biometric databases or to enhance the recognition accuracy of the various primary biometric traits.

In the recent past, there has been increased demand for personal identification and verification of authorized users of automated systems. Biometrics offers an effective approach to establish personal identity by using individual's unique, reliable and stable physical or behavioral characteristics. Biometrics is concerned with the unique, reliable and stable personal physiological characteristics such as fingerprints, palmprint, facial features, iris pattern, retina, and hand geometry. An efficient and effective automated biometric system is required for authentication and identification of an individual's gender. One of the oldest and most basic examples of the characteristics that are used for recognition by humans is the face. In this paper, we have presented a scheme based on multiresolution analysis which uses discrete wavelet transform for gender identification using frontal faces.

1.1 Related Work

Gender identification is a well-studied problem in the domain of biometric technologies; some of the earlier works related to gender identification are briefly presented in this paragraph. In [1], the authors studied gray level co-occurrence matrix and discrete wavelet transform-based method to extract features from frontal facial images. SVM classifier with tenfold cross validation is used for classification task and they achieved gender classification accuracy of 88%. A comparative study on large datasets of wild faces is carried out in [2], Histogram of oriented gradients and local binary patterns are compared with Convolution Neural Networks. In experimental analysis, the authors found that score level fusion would be a suitable solution for the problem of gender classification in wild and cross databases. Deep learning-based method is presented in [3] with built-in feature selection based on Linear Discriminant Analysis and an accuracy of 98% was achieved on LFW dataset. Gabor filters and Binary patterns based combined approach is presented in [4]. Using SOM Neural Networks, the authors achieved an accuracy of 90.33% in the task of gender identification. In [5], the authors presented comparative analysis of different machine learning algorithms for gender identification using frontal faces. Fuzzy inference-based technique is developed in [6] for gender identification in unconstrained cross database of facial image. Independent component analysis-based approach is presented in [7] to identify gender from frontal facial images from FERET dataset. Using SVM with ICA space, they achieved the accuracy of 96%. Discrete Wavelet transform-based features are used in [8] with the combination of SVM classifiers. In [9], hybrid features based

Input face Image
↓
$\boxed{\text{Preprocessing}} \Rightarrow \boxed{\text{Features Computation}} \Rightarrow \boxed{\text{Classification}}$
↓
Predicted Gender Output.

Fig. 1 Schematic diagram of our method

on Discrete Wavelet transform and LBP were presented for gender identification using facial images. From the above brief review, we observed that global features have shown promising results in gender identification from facial images. But huge dimensionality is one of the problems with global descriptors. In this paper, we have presented a global method based on histogram of Discrete Wavelet Transform-based coefficients and their statistical measures. This reduces high dimensional coefficients into fixed size features vector as compact representation of input facial image.

2 Proposed Method

Our method consists of three main steps preprocessing, feature extraction, and classification. In preprocessing step, we have performed image enhancement operation; to extract features Discrete Wavelet transform is employed and the task of classification is studied with three classifiers, namely Nearest Neighbor, Support Vector Machines, and Linear Discriminant Analysis. The block diagram of our proposed method is given in Fig. 1.

2.1 Preprocessing

Facial images in our dataset are well aligned and resized, and therefore in preprocessing, we have simply attempted image normalization operation, to balance the varying lighting conditions. First input image is converted from RGB to grayscale, where pixels are represented in two-dimensional space ranging from 0 to 255 gray levels. To normalize the grayscale, face image Contrast Limited Adaptive Histogram Equalization (CLAHE) was used. The advantage of this method over traditional histogram equalization is that it provides enhancement of contrast at the part of the image wherever necessary. More details on CLAHE are given [10]. For better understanding, we have shown preprocessing in Fig. 2.

Fig. 2 An example showing preprocessing

2.2 Feature Extraction

Representing essential discriminative information is one of the goals of feature extraction techniques in digital image processing applications. In this paper, our main focus is on feature extraction from frontal facial images to obtain the gender information. To do this, we have used Discrete Wavelet Transform (DWT) [11], a well-known frequency domain method for Feature Extraction in this application. DWT is widely used in Image Compression [12] Filtering [13] and Feature Extraction [14]. It decomposes input image signal into different directions by analyzing at various scales and directions. Discrete Wavelet Transform in two dimensions decomposes signal into detail and approximation subbands. In approximation subband, the overall information of an input image is preserved and horizontal, vertical, diagonal details are presented in the detail subbands. To extract fixed size feature vector from facial images, first discrete wavelet Transform is applied up to four levels to decompose image into 16 DWT subbands and only 12 detail subbands are considered for feature computation, whereas 4 approximation subbands are excluded. For better visualization, only two-level decomposition of female face using DWT is shown in Fig. 3. The process of feature computation is explained briefly in the paragraph below.

Let C_i be the discrete random variable that denotes coefficients values in DWT subband and $h(C_i)$, $i = [0, 1, 2, 3, \ldots, k-1]$ be the normalized histogram, where k is the possible DWT subband coefficient value. Shape analysis of computed histogram is performed via its central moment and the expression of the nth moment for mean can be expressed as follows:

$$\mu_i = \sum_{a=1}^{k} (C_i - m)^n \tag{1}$$

From the abovementioned central moment, statistical descriptors such as mean (F1), contrast (F2), relative smoothness (F3), skewness (F4), uniformity (F5), and entropy

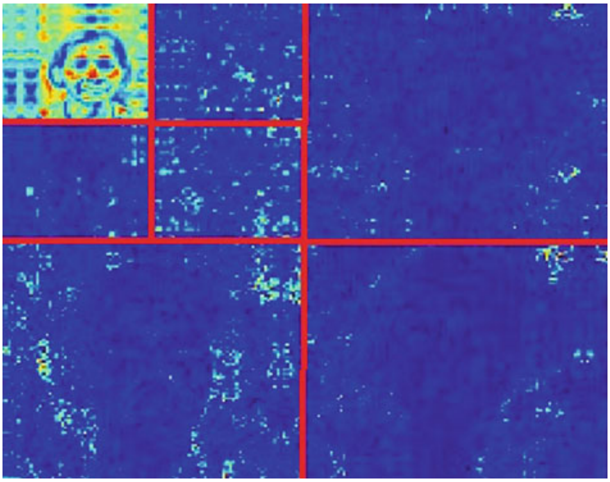


Fig. 3 Visualization of DWT decomposition of face image

(F6) are computed. In addition to this, each subband is further convolved with local standard deviation filter, local entropy filter and local range filter, and then from each filtered subband as single statistical value standard deviation is computed. From this procedure, we get three more features each from one filtered subband (F7, F8, and F9), more details on this are given in [15, 16]. In this way, fixed size feature vector of four levels $\times 3$ subbands $\times 9$ features = 108 features is computed from each input facial image.

Algorithm: Computation of Multiresolution Statistical Descriptors.
Input: Frontal Face Image F
Output: 108 dimension Features Vector

---

1. Compute DWT of (F)
2. Extract Vertical, Horizontal, and Diagonal DWT sub bands of input face image by performing four level decomposition.
3. At each level, for each subband Compute Histogram of DWT subband coefficients.
4. Compute mean (F1), contrast (F2), relative smoothness (F3), skewness (F4), uniformity (F5), and entropy (F6), these six statistical texture descriptors from each DWT subband histogram.

5. Apply local standard deviation filter, local range filter, and local entropy filter to each sub band separately.

6. For each filtered subband, compute standard deviation (three features each from one filter).

7. Concatenate features from step 4 and step 6 to form feature vector. We get three sub bands * 9 texture features * 4 level = 108 Features.

2.3 Classification

Nearest Neighbor Classifier: Nearest Neighbor, one of the most popular, basic, and simple classifier, is considered for the classification. One of the important things is that, it does not require actual training, whereas requirement of memory increases with training samples. Based on user-defined suitable distance measure it searches for nearest neighbor and gives label to unlabeled sample. To search nearest neighbor in this work, we have used Euclidean distance.

Let $A(a_1, a_2, a_3 \ldots, a_n)$ be the training sample and $B(b_1, b_2, b_3, \ldots, b_n)$ be the testing samples. Euclidean distance between A and B is defined as follows:

$$d(A, B) = \sqrt{(a_i - b_i)^2} \tag{2}$$

Support Vector Machine: Support Vector Machine (SVM) is statistical learning based classifier invented by Vapnick [17] SVM tries to find out an optimal hyperplane which separates the two classes for a set of n data vectors say X_i. A descriminant function $g(X) = W^T \cdot X - b$ separates each data item into two classes

$$g(X) = W^T \cdot X_i - b \geq 1 \tag{3}$$

where y_i is the class either $+1$ (male) or -1 (female) in our case.

Linear Discriminant Analysis: Linear Discriminant Analysis (LDA) is one of the most commonly used techniques, because of its simplicity, generalization capacity, and easy computation. The main idea behind the classification in LDA is to enhance the separation between classes by maximizing the ration of between classes to within-class variance. Here, LDA tries to identify male and female class of facial image based on between and within class variance, more details are give in [18].

3 Experiments

3.1 Dataset and Evaluation Protocol

Well-known publicly available dataset of frontal facial images available from [19] is considered for experiments. In the present study, a subset of 840 images (420 male and 420 female) is used for experiment purpose. These images were captured with varying pose, expressions, and lighting conditions. The images from dataset are not shown in the paper, due to privacy policy and instructions provided on dataset web page. For illustration purpose, we have used the face image of one of the volunteer from our Lab (Department of Computer Science, Karnatak Arts, Science and Commerce College, Bidar). Informed consent is also collected from the volunteer about the use of her photograph in research article.

To evaluate the performance of the method, we have used tenfold cross validation. First, the complete data is divided into ten subparts, when any one subpart serves as testing set, then other nine subparts are used as training set. This procedure is repeated ten times in such a way that each subpart gets an opportunity to serve in training and testing process. The average of these trials will be considered as final result. In this study, we have defined the accuracy as follows:

$$Accuracy = \frac{\text{\#No. of faces correctly classified in a class}}{\text{\#Total no. of faces in a class}} \times 100 \qquad (4)$$

3.2 Results and Discussion

To investigate the gender discriminating ability of multiresolution features in frontal facial images, we employed three classifiers, namely Nearest Neighbor Classifier, Support Vector Machine, and Linear Discriminant Analysis. The results obtained during the experiments are shown in Table 1.

From Table 1, it can be observed that Nearest Neighbor Classifier outperformed in the task of gender identification using frontal faces with multiresolution statistical descriptors as compared with LDA and SVM and yields an accuracy of 96.42% for

Table 1 Gender identification accuracy in % using multiresolution statistical descriptors

Classifier	Male	Female	Overall accuracy
Nearest neighbor	96.42	97.85	97.14
Support vector machine	91.42	94.76	93.10
Linear discriminant analysis	88.80	96.90	92.85

Table 2 Confusion matrices for all three classifier

Nearest neighbor	
Male	Female
405	15
9	411
SVM	
Male	Female
384	22
36	398
LDA	
Male	Female
373	47
13	407

male, 97.85% for female and overall accuracy of 97.14%. LDA gave poor performance with the accuracy of 88.80% for male, 96.90% for female and overall accuracy of 92.85%. SVM served in between Nearest Neighbor and LDA with the accuracy of 91.42% for male, 94.76% for female and overall accuracy of 93.10%. For deeper understanding, we have given confusion matrices for all three classifiers in Table 2. From the experimental results, gender discriminating ability of the presented multiresolution statistical descriptors from frontal facial images is proved. The obtained accuracy of 97.14% is quite promising with complex dataset of frontal facial images. The main aim of the experiments presented in this paper is to prove the significance of compact statistical descriptors obtained using the histogram of DWT subbands in the task of gender identification based on frontal facial images. Comparison with state-of-the-art techniques is given in Table 3. From the facts and figures given in

Table 3 Comparison with existing approaches and proposed work

Methods	Dataset	Features	Accuracy (%)
Ergen et al. [1]	FEI	GLCM and DWT	87
JafariBarani et al. [4]	AR Face	Gabor and LBP	90
Lemley et al. [5]	FERET	CNN	96
TanerDanisman et al. [6]	FERET	ICA	95.67
Walczak et al. [8]	FERET	DWT	90
Proposed	Collection of facial images from university of essex	DWT based multiresolution statistical descriptors	97.14

Table 3. It can be noted that, our method outperformed in the task of gender identification as compared to existing one. In the future, we will explore the investigation of proposed features for gender identification using cross databases.

4 Conclusion

In this paper, we have presented multiresolution statistical descriptors for gender identification using frontal faces. Multiresolution statistical descriptors are computed by analyzing the shape of histogram obtained from DWT subbands. We achieved the promising accuracy of 97.14% with very basic and simple Nearest Neighbor classifier. We proved that the presented features have strong discriminating capacity and they can be applied effectively for the problem of gender determination using frontal facial images. In future, we will evaluate the proposed descriptors with publicly available large dataset of facial images.

Acknowledgements We are thankful to Smt. Savitri A. Nawade for participation in the creation of database for the experimentation work stated in this paper.

References

1. Ergen, B., Abut, S.: Gender recognition using facial images. In: proceedings of International Conference on Agriculture and Biotechnology IPCBEE, vol. 60. no. 22, pp. 112–117. IACSIT Press, Singapore (2013)
2. Castrilln-Santana, M., Lorenzo-Navarro, J., Ramn-Balmaseda, E.: Descriptors and regions of interest fusion for in- and cross-database gender classification in the wild. Image Vis. Comput. **57**, C, 15–24 (2017)
3. Tian, Q., Arbel, T., Clark, J.J.: Deep LDA-Pruned nets for efficient facial gender classification. In: CVPR Workshops, pp. 512–521 (2017)
4. JafariBarani, M., Faez, K., Jalili, F.: Implementation of gabor filters combined with binary features for gender recognition. Int. J. Electr. Comput. Eng. **4**(1), 108–115 (2014)
5. Lemley, J., Abdul-Wahid, S., Banik, D., Andonie, R.: Comparison of recent machine learning techniques for gender recognition from facial images. In: Proceedings of the 27th Modern Artificial Intelligence and Cognitive Science Conference 2016, Dayton, OH, USA, pp. 97–102, 22–23 Apr 2016
6. TanerDanisman, I.M., Bilasco, J.M.: Boosting gender recognition performance with a fuzzy inference system. Expert Syst. Appl. **42**(5), 2772–2784 (2015)
7. Jain, A., Huang, J., Fang, S.: Gender identification using frontal facial images. In: 2005 IEEE International Conference on Multimedia and Expo, 4 p (2005)
8. Walczak, J., Wojciechowski, A.: Improved gender classification using discrete wavelet transform and hybrid support vector machine. Mach. Graph. Vis. **25**(1/4), 27–34 (2016)
9. Hussain, M., Ullah, I., Aboalsamh, H.A., Muhammad, G., Bebis, G., Mirza, A.M.: Gender recognition from face images with dyadic wavelet transform and local binary pattern. Int. J. Artif. Intell. Tools **22**(06), 1360018 (2013)
10. Pizer, S.M., Amburn, E.P., Austin, J.D., et al.: Adaptive histogram equalization and its variations. Comput. Vis. Graph. Image Process. **39**, 355–368 (1987)

11. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**(7), 674–693 (1989)
12. Singh, V.K.: Discrete wavelet transform based image compression. Int. J. Remote Sens. **20**(17), 3399–3405 (1999)
13. Demirel, H., Anbarjafari, G.: Discrete wavelet transform-based satellite image resolution enhancement. IEEE Trans. Geosci. Remote Sens. **49**(6), 1997–2004 (2011)
14. Hiremath, P.S., Pujari, J.D., Shivashankar, S., Mouneswara, V.: Script identification in a handwritten document image using texture features. In: IEEE 2nd International Advance Computing Conference (IACC), Patiala, vol. 2010, pp. 110–114 (2010)
15. Pardeshi, R., Chaudhuri, B.B., Hangarge, M., Santosh, K.C.: Automatic handwritten Indian scripts identification. In: 2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 375–380. IEEE, Sept 2014
16. Hangarge, M., Santosh, K.C., Doddamani, S., Pardeshi, R.: Statistical texture features based handwritten and printed text classification in south Indian documents (2013). arXiv:1303.3087
17. Vapnik, V.: The Nature of Statistical Learning Theory. Springer Science & Business Media (2013)
18. Uray, M., Roth, P.M., Bischof, H.: Efficient classification for large-scale problems by multiple lda subspaces. In: Proceedings of International Conference on Computer Vision Theory and Applications, pp. 299–306 (2009)
19. Face Recognition Data. http://cswww.essex.ac.uk/mv/allfaces/index.html

Captioning the Images: A Deep Analysis

Chaitrali P. Chaudhari and Satish Devane

Abstract Image captioning is one of the fundamental tasks in machine learning since the ability to generate text captions of an image can have a great impact by assisting us in day-to-day life. However, it is not just an object classification or recognition task, because the model must know the dependencies among the recognized objects and their attributes and encode that knowledge correctly in the caption using a natural language like English. Recently, the internet is overwhelmed with the huge amount of textual and visual data consisting of billions of unstructured images and videos. Meaningful captions will serve as useful keys for retrieval, creative searching, and powerful browsing of these images. In this paper, we present the goal of analysis and classification of the recent state-of-the-art in image captioning and discuss significant differences among them. We provide a comparative review of existing models, techniques with their advantages and disadvantages. Future directions in the field of automatic image caption generation are also explored.

Keywords Image captioning · Natural language processing · Computer vision

1 Introduction

The text or images individually may give ambiguous information but jointly they convey a meaningful description. Manually, annotating images is uncertain, time—consuming, and error-prone and hence nearly impossible. Thus, it is important and interesting to analyze ways to automatically describe the contents of an image using natural language. The trademark characteristic of us to describe images in great details by just having a quick glance at it is a very difficult task for the image caption generation models. To improve the precision and context in image captioning, a

C. P. Chaudhari (✉)
Lokmanya Tilak College of Engineering, Koparkhairane, Navi Mumbai, India
e-mail: chaitralichaudhari13@gmail.com

S. Devane
Datta Meghe College of Engineering, Airoli, Navi Mumbai, India

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_100

step forward can be taken by considering the semantics of an image for basic understanding and to lead captioning beyond the tasks of image classification and object recognition. In addition, to encourage the success of complex models, computational power should be enhanced. Image understanding makes use of pattern recognition mechanism. Intermediate representation of original image is generated and then object identification and classification tasks are carried out. For identification and classification of objects, the three-dimensional descriptions of the objects and a particular knowledge about the scene are extracted from the intermediate description. Higher level information is used to simplify the problem and make the processing at lower level easier.

Automated image captioning still is not perfect, and it has quickly become a hot research area. We surveyed the existing literature available and presented an overview of state-of-the-art techniques and models. The rest of the paper is organized as follows. Section 2 introduces Image Captioning, Sect. 3 briefly describe some recent research works in this area with their benefits and limitations, Sect. 4 presents summarized findings based on the literature survey, Sect. 5 introduces comparative analysis based on survey, and Sect. 6 introduces the conclusion and ideas for future work.

2 Image Captioning

Automatically describing the contents of an image in a meaningful, self-contained, grammatically, and semantically correct way using a natural language is termed as image captioning. It can be placed in the context of the broader question of Image Understanding. New research opportunities in image captioning are explored with the application of deep neural networks which has shown significant success in semantic understanding of images. The image captioning systems tells us what the image is all about. Humans are having the ability to quickly identify the situation in a complex scene and summarize it instantly in understandable words which is difficult for a computer. The prominent features are considered by the humans ignoring the unimportant details. For instance, if we see benches and a blackboard, we identify the image is of a classroom.

2.1 Areas of Application

Image captioning can impart artificial intelligence to machines allowing them to analyze and describe complex visual scenes. Captions provide information, e.g., nouns which can help train object detectors, written descriptors of objects can train object recognizers, and adjectives can be used to train material recognizers.

Other applications include generating summaries for videos, listening to a lecture, understanding slides, automatic video surveillance, describing pictorial contents to the visually impaired, retrieval of image data by search engines to respond user queries, news caption generation, medical image captioning, education and learning, security, military applications as well as a tour guide. Algorithms are developed to address the needs of more practical and recent developments of vision computing like the use of wearable cameras and lightweight computing devices. The context of the content from the images/videos can be used by a visually challenged person to interact with the external world. For example, to read a text or sign, locate a specific object, pose of the object, and identify a person with his facial expressions.

2.2 Existing Scenario

Ideally, the text generated by image captioning system is expected to describe what is visually depicted in the image using a single sentence. Despite the difficult nature of this task, computer vision, and natural language processing researchers have made significant strides in this area. But the ability of these models to understand, analyze, and describe complex images, is still an active area of research. Dominant models are making the use of deep learning networks. Standard, good quality, and freely available datasets are available to train these networks. Success of the methods depends a lot on richly annotated images, acquiring which is both time-consuming and expensive. The sources of weakly tied images like ImageNets are available. However, a really high-quality data is required while most of the data available in recent datasets is noisy.

2.3 Future and Expectation

Machines are to be trained to identify the salient features in the image, extract the actual contextual information, and push away the unimportant details in the image. Future work involves improving the performance of existing models for image captioning to be used for other multimodal applications such as video description, speech to text conversion, language translation, and determining their efficiency. Exploration of lower CNN layers for initialization and learning representation of characters, and study of different RNN sizes is required. For solving the image captioning problem using low or limited data "visual recognition under weak supervision" approach is promising. There are other ways of using previous knowledge or knowledge in other domains with learning techniques like "transfer learning", "multitask learning" and "semi supervised learning".

Applications have been emerged such as visual question answering. For example, the image is captioned as: A girl playing a ball and a question could be posed with that image as: What color ball is the girl playing? Giving a possible answer as "Red". These problem areas are interesting because they combine the advantages

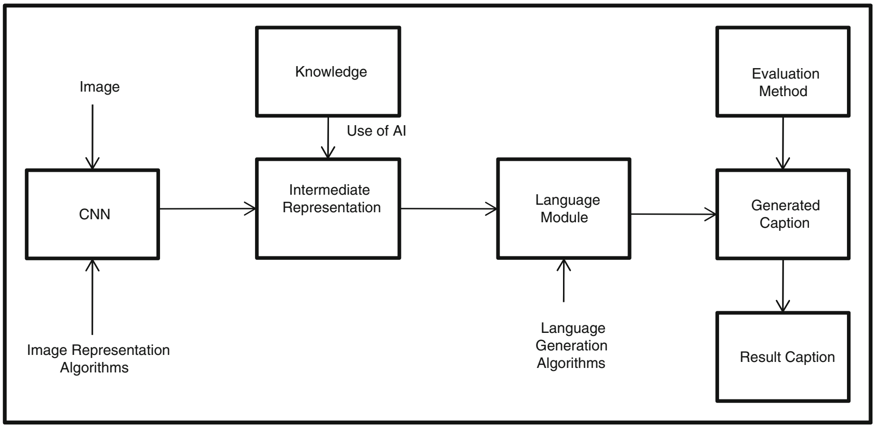


Fig. 1 A general framework for image caption using neural network

of computer vision and natural language models and algorithms to address relevant challenges and applications.

2.4 Framework Representation

The general framework for image captioning consists of three major parts: gathering images of complex everyday scenes containing common objects in their natural context, selection and improvisation of the image representation and captioning algorithms/techniques, and evaluation of the generated captions (Fig. 1).

2.5 Classification of Approaches

The image captioning models the following bag-of-words/features approach treat words/features as floating tags and look at the co-occurrence of image regions and annotated words. These are generative models following the hierarchical structure which encourages the better capturing of semantics through levels of generalization. This image representation may be inputted to the NLG model to describe images. The second category contains image captioning model based on classifier approach. The third category is image captioning model based on retrieval approach which evaluates the mapping between images and sentences independent of the generation aspect by associating an image with a description and then retrieving and ranking a set of similar images with candidate descriptions. The retrieval of images and ranking of their descriptions can be carried out in two ways: either using perception or from a multimodal space that combine textual and visual information space. Lastly, image

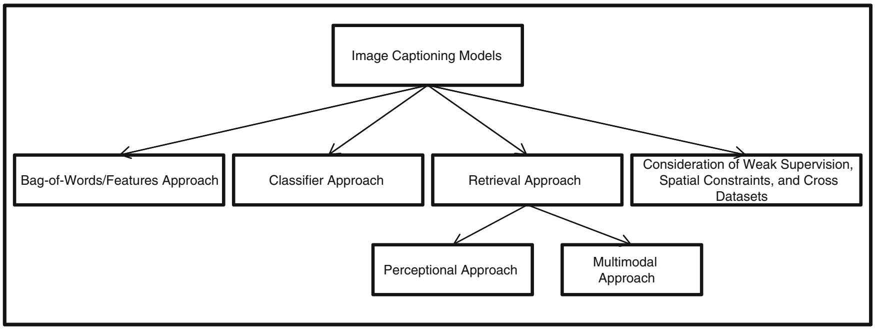


Fig. 2 Image captioning models categorization

description models are categorized using spatial constraints, weak supervision and cross-datasets (Fig. 2).

3 Review of Literature

Automatic image caption generation involves two major tasks: analysis of image (image representation) and generation of textual description. Image representation includes the tasks like identification of objects, their attributes, image features, and interaction between the objects. This representation is expressed in concise and grammatically correct sentences by Natural Language Generator (NLG) model. We extensively surveyed the literature and categorized the image captioning models based on their approaches towards the description generation.

3.1 Image Captioning Models Based on Bag-of-Words/Features Approach

We take a glance towards these models focusing on the two major aspects, i.e., Image representation and Sentence generation (Natural Language Descriptions). Fischler and Elsclager [1] presented a model to find the object in an actual photograph, given some description of a visual object. The desirable feature of the approach was its generality, and the model was called as an embedding model. Barnard and Forsyth [2], Barnard et al. [3, 4], focused on ways of learning the links between image features and associated text (semantics). The model for object recognition by Duygulu et al. [5] was described as a machine translation model where K means is used to vector quantize the image region representation. The label associated with a region with this process is called as a blob. Whereas, Barnard and Forsyth [2], Barnard et al. [3, 4] presented the models as the statistical models inspired by hierarchical mixture model given by Hofmann in [6].

Farhadi et al. [7] proposed a model with an intermediate space of meaning which has different projections to the space of images and space of sentences; Kulkarni et al. [8] have used projections between objects, attributes as well as propositions; whereas, Yang et al. [9] have presented a language model trained from the English Gigaword corpus to obtain their projections. With the aim to describe the world represented in an image in a human manner, Li et al. [8] have represented an image as a set of triples with each triple encoding a spatial relation between two objects; Yatskar et al. [10] used densely annotated images with human-generated sentence descriptions and Fang et al. [11] used a caption generator trained from a dataset of images. To encode the geometric relations between the regions of an image, Elliott and Keller [12] have introduced visual dependency representations (VDRs) to represent the structure of images and later with a slight modification, Elliott and de Vries [13] used (VDRs) with object detectors and generated image representations. Ortiz et al. [14] and Lin et al. [15] have chosen the problem of interpreting and verbalizing visual information using abstract and indoor scene, respectively.

Methods of generating captions of visual content can be divided into two main types of approaches: methods that make use of existing text to caption images and methods that predict image content and then build captions from scratch.

Duygulu et al. [5] used the EM algorithm to find the correspondence between blobs and words; Li et al. [16] have proposed to compose fresh sentences with the approach consisting of two steps: (n-g) phrase selection and (n-gram) phrase fusion; Kulkarni et al. [8] followed the similar model, and generated captions by filling sentence template slots with words selected from a conditional random field that predicts the most likely image labeling, whereas the Fang et al. [11] used a maximum entropy language model to generate image captions. All these models generated captions from scratch. Some other methods use sentence templates to generate captions. Yang et al. [9] used HMM that models the sentence generation process, with hidden nodes as sentence components and image detections as the emissions whereas Elliott and Keller [12] used visual dependency representations in template-based description generation models. Some more approaches are linguistically sophisticated towards generation, e.g., Kuznetsova et al. [17] addressed image caption generalization as an optional subtask of image caption generation; Ortiz et al. [14] converted the abstract scenes dataset into a parallel corpus of visual and linguistic descriptions, which allowed to train a statistical machine translation (SMT) model; Mitchell et al. [18] used the system ("Midge") to generate a well-formed description of an image in present tense, declarative phrases, as a naive viewer without prior knowledge of the photograph's content.

In Summary, these approaches were quite successful in describing images, but they are heavily hand designed. Also, their text generation power is fixated on the database/embeddings and is not able to describe previously unseen compositions of objects.

3.2 Image Description Models Based on Classifier Approach

This was carried out for various restricted domains such as identifying images containing cars as proposed by Fergus et al. [19], where objects are represented as flexible constellations of parts and the model was called as constellation model; or providing powerful cues as to who is depicted in the associated news image as in Berg et al. [20]; even for a dataset of inaccurately and ambiguously labeled face images in the news Berg et al. [21]; grids of locally normalized Histograms of Oriented Gradient (HOG) descriptors for human detection used by Dalal and Triggs [22]; categories of animals as in Berg et al. [23], Lampert et al. [24, 25]; identification of faces as in Kumar et al. [26, 27]. These works had focused on use of text as a source for predicting the image content, i.e., associating names in the captions to objects in the images. By modeling both names and action verbs jointly generalization was attempted by Jie et al. [28].

The models with iterative incremental learning framework by Li [29] and a multimodal approach by Schroff [30] were proposed to automatically collect larger diverse object datasets from the web with ample intra-class variation and to incrementally learn object category models. Lampert et al. [24, 25] proposed a probabilistic model to solve the problem of learning with disjoint training and test classes using attribute-based classification. An approach considering the problem of detecting and localizing objects of a generic category used a multiscale deformable part model by Felzenszwalb [31, 32].

The image description models the following classifier approach to solve the problem as a classification problem. Inputted images are compared against the images in the databases. The textual representation associated with that image is reused or reformed to generate the novel description and/or object classification.

3.3 Retrieval Using Perception

Olivia et al. [33] proposed a computational model of the recognition of real-world scenes based on a very low-dimensional representation of the scene, which is termed as the Spatial Envelope. Ordonez et al. [33] used a large captioned photo collection associated with visually descriptive text and done description generation using relatively simple nonparametric methods. Kuznetsova et al. [34] identify content elements present using the classifiers and detectors and then retrieve phrases referring to those content elements from the database by a generation process cast as a constraint optimization problem. Mason and Charniak [35] used a nonparametric density estimation technique for image caption generation by effectively leveraging the information from the massive amounts of human-written image captions on the internet. Patterson et al. [36] build the “SUN attribute database” using crowd sourcing to annotate attributes for 14,340 images from 707 scene categories. Yagcioglu et al. [37] used averaging the sentence vectors extracted from the captions of images similar to the input image and translated a visual query into a distributional semantics form. Devlin et al. [38] combined key aspects of the ME and RNN methods and achieve a new record performance.

Retrieval models using perception typically require a large amount of training data to generate consistent captions of the images.

3.4 Retrieval Using Multimodal Approach

Given a set of pictures and a few word queries; the retrieval system outputs a picture ranking in which the pictures relevant to the query appear above the others. Grangier and Bengio [39] have proposed a model which addresses the retrieval problem directly and does not rely on an intermediate image annotation task. Hodosh et al. [40] have proposed evaluating models by scoring a pool of unseen captions, using Kernel Canonical Correlation Analysis (KCCA). Socher et al. [41] used dependency trees to embed sentences into a vector and retrieve images described by those sentences by dependency tree RNN (DT-RNN) model. Karpathy et al. [42] used a deep, multimodal embedding of visual and natural language data for bidirectional retrieval of images and sentences. Sun et al. [43] have used a parallel text and visual corpora and suggested an automatic visual concept discovery algorithm which filtered the text terms based on the visual discriminative power of the associated images, and grouped them into concepts using visual and semantic similarities. Pinheiro et al. [44] presented a simple bilinear model able to generate descriptive sentences given a sample image by learning a metric between an image representation and phrases that are used to describe them. This model has strongly focused on the syntax of the descriptions. Ushiku et al. [45] proposed a phrase-learning method called as Common Subspace for Model and Similarity (CoSMoS). For sentence generation, a beam-search-based decoder or templates are used. Kiros et al. [46] introduced an encoder–decoder pipeline approach that unifies joint image text embedding models with multimodal neural language models. Donahue et al. [47] described end-to-end trainable class of recurrent convolutional architectures suitable for large-scale visual understanding tasks. Karpathy and Fei-Fei [48], proposed a novel multimodal Recurrent Neural Network architecture which takes an input image and generates its description in text. Xu et al. [49] proposed an attention-based model using standard backpropagation techniques that automatically learns to describe the content of images. An attention component determines which regions in an image are salient, and it can focus its description on those regions. Lebret et al. [50] proposed a purely bilinear model with a strong focus on the syntax of the descriptions that learns a metric between an image representation and phrases that are used to describe them.

The neural image caption generator framework is trained on large numbers of image caption pairs, and the model learns to capture relevant semantic information from visual features which can be used to generate novel image descriptions. Recurrent Neural Network (RNN) architectures include the Simple Recursive Networks

(SRNs) and Long Short-Term Memories (LSTMs) which enable solutions to basic problems such as activity recognition and object trajectory prediction.

3.5 Image Description Models Using Spatial Constraints, Weak Supervision and Cross-Datasets

Gupta et al. [51] and Oliva et al. [52] presented an approach that applies spatial and functional constraints on each of the perceptual elements for coherent semantic interpretation to recognize objects and actions from static images without using any motion information when the appearances are not discriminative enough. Lazebnik et al. [53] presented a method for recognizing scene categories based on a "spatial pyramid" which showed a significantly improved performance on challenging scene categorization tasks. Where most of the models require a training set of fully annotated image sentence pairs which is prohibitively expensive, Prest et al. [54] used weak supervision approach that transfers knowledge from millions of weakly annotated images to improve the accuracy of description retrieval. Verma et al. [55] proposed a Structural SVM based unified formulation for two complementary cross-modal prediction tasks: (i) predicting text(s) given an image ("Im2Text") and (ii) predicting image(s) given a piece of text ("Text2Im"). They highlight the importance of cross-dataset image description retrieval evaluation.

4 Summarized Findings of the Literature Review

Uncertainties about salient content, main subject detection, object recognition, action detection, and scene understanding make image captioning a challenging problem. Inaccuracies in the description vary from minor errors that can be ignored to image descriptions that do not go very well with the test image. Most common mistakes include: singular versus plural mistakes in the descriptions; presence of words that are commonly associated with one another although they are not present in the image; inaccurate action recognition due to the absence of visual temporal information; and failure to identify unforeseen objects or unforeseen views of an object leading to inaccurate descriptions.

5 Comparative Analysis

Bernardi et al. [56], classified existing approaches based on how they conceptualize the problem. This paper presented the comparative review of existing models and techniques with their advantages and disadvantages. Compared to the traditional captioning approaches, the recent approaches show significant improvements to the computer vision component which is much faster to train; produce more detailed and accurate object detection, image classification and description generation. But they suffer from the drawback of vast information requirement about the world in the form of images, videos and text collected and used as datasets. Also, the existing object recognition systems struggle to recognize objects in the background, obstructed objects or disordered objects which actually represent the composition of everyday scenes and can be identified only using the context information. Lenat et al. [57] have created CYC to assemble a standard ontology and common knowledge base in a formal language which involves handling causality, time, space, substances, intention, contradiction, uncertainty, belief, emotions, planning, and so on which is worth mentioning here. In current research, image captioning models are trained from supervised training data where images are annotated by hand with multiple very descriptive sentences, sometimes also localized in the image. These techniques are useful for preliminary research, but not suitable where datasets are large and diverse. Generating annotations for such data sets is too costly.

Secondly, the earlier strategy of considering pixels to identify object has a main drawback that it loses the information of an object on the contrary, by adopting the strategy of considering segments instead of pixels, as long as the segmentation module can produce object-consistent regions, the output will naturally be an object-consistent result. Attention-based models use the CNN feature extracted from higher layers, which increases the computational cost. Also as the amount of training data increases, the model with less training time will be needed. Effective architectures are now very deep or ultra-deep and need very long training times even up to several weeks. It is not efficient to retrain a complete model for several new categories or for applying already trained categories to different domains.

6 Conclusion

In conclusion, the ways to represent the common cases of various phenomena, to reason about them efficiently, and to formulate sets of categories and attributes are required to explore. Transferring knowledge between object classes is required which may reuse the information and enable object recognition with scarce training data. The need to device effective algorithms to process weakly annotated images for caption generation is identified which will better withstand the real-life applications. Improvement to the vision and language component is required so that the captioning model can more specifically address the problem of description generation than object classification. For further progress the areas shall have to get researched in depth like how to design the architectures to model abstract concepts and theories, how to acquire, store and manipulate it in a framework, how to represent concepts and set the objectives to make artificial agents understand the world from human perspective. Also, insights to better generalization capabilities of the model should be provided. Better feature localization in the image can greatly improve the performance of retrieval tasks, and similar improvement might be seen in the description generation task. Region-based encoder–decoder models for images, however, are promising because they improve generalization; enable relative geometrical statements and grounding of properties and attributes to individual object instances. With this, we see the caption generation systems have a long way to go towards generation of meaningful image captions.

References

1. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Trans. Comput. **22**(1) (1973)
2. Barnard, K., Forsyth, D.: Learning the semantics of words and pictures. In: Proceedings of International Conference on Computer Vision, pp. II:408–415 (2001)
3. Barnard, K., Duyguly, P., Forsyth, D.: Clustering art. In: Proceedings of IEEE Conference Computer Vision and Pattern Recognition, June 2001
4. Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., Jordan, M.: Matching words and pictures. J. Mach. Learn. Res. **3**, 1107–1135 (2003)
5. Duygulu, P., Barnard, K., de Freitas, N., Forsyth, D.: Object Recognition as machine Translation: Learning a lexicon for a fixed image vocabulary. In: Proceedings of European Conference Computer Vision (2002)
6. Hofmann, T.: Learning and representing topic: a hierarchical mixture model for word occurrence in document databases. In: Proceedings of Workshop on Learning from Text and the Web, CMU (1998)
7. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: generating sentences from images. In: European Conference on Computer Vision (2010)
8. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A., Berg, T.: Baby talk: understanding and generating simple image descriptions. In: Computer Vision and Pattern Recognition (CVPR), vol. 2011, pp. 1601–1608. IEEE (2011)
9. Yang, Y., Teo, C.L., Daume, III.H., Aloimonos, Y.: Corpus-guided sentence generation of natural images. In: Conference on Empirical Methods in Natural Language Processing (2011)
10. Yatskar, M., Galley, M., Vanderwende, L., Zettlemoyer, L.: See no Evil, say no Evil: description generation from densely labeled images. In: Joint Conference on Lexical and Computation Semantics (2014)
11. Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Doll´ar, P., Gao, J., He, X., Mitchell, M., Platt, J., Zitnick, C.L., Zweig, G.: From captions to visual concepts and back. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
12. Elliott, D., Keller, F.: Image description using visual dependency representations. In Conference on Empirical Methods in Natural Language Processing (2013)
13. Elliott, D., de Vries, A.P.: Describing images using inferred visual dependency representations. In: Annual Meeting of the Association for Computational Linguistics (2015)
14. Ortiz, L.M.G., Wolff, C., Lapata, M.: Learning to interpret and describe abstract scenes. In: Conference of the North American Chapter of the Association of Computational Linguistics (2015)
15. Lin, D., Fidler, S., Kong, C., Urtasun, R.: Generating multi-sentence natural language descriptions of indoor scenes. In: British Machine Vision Conference (2015)
16. Li, S., Kulkarni, G., Berg, T., Berg, A., Choi, Y.: Composing simple image descriptions using web-scale n-grams. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pp. 220–228. Association for Computational Linguistics (2011)
17. Kuznetsova, P., Ordonezz, V., Berg, T.L., Choi, Y.: TREETALK: composition and compression of trees for image descriptions. In: Conference on Empirical Methods in Natural Language Processing (2014)

18. Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A. C., Yamaguchi, K., Berg, T. L., Stratos, K., Daume, III.H.: Midge: generating image descriptions from computer vision detections. In: Conference of the European Chapter of the Association for Computational Linguistics (2012)
19. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: CVPR (2003)
20. Berg, T.L., Berg, A.C., Edwards, J., Forsyth, D.A.: Who's in the Picture? In: Proceedings Neural Information Processing Systems Conference (2004)
21. Berg, T.L., Berg, A.C., Edwards, J., Maire, M., White, R., Learned-Miller, E., Teh, Y.-W., Forsyth, D.A.: Names and faces in the news. In: Proceedings of IEEE Conference Computer Vision and Pattern Recognition (2004)
22. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
23. Berg, T.L., Forsyth, D.A.: Animals on the web. In: Proceedings of IEEE Conference Computer Vision and Pattern Recognition (2006)
24. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
25. Lampert, C., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
26. Kumar, N., Belhumeur, P., Nayar, S.K.: FaceTracer: a search engine for large collections of images with faces. In: ECCV (2008)
27. Kumar, N., Berg, A.C., Belhumeur, P., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV (2009)
28. Jie, L., Caputo, B., Ferrari, V.: Who's doing what: joint modeling of names and verbs for simultaneous face and pose annotation. In: NIPS, editor, Advances in Neural Information Processing Systems, NIPS (2009)
29. Li, L.-J., Fei-Fei, L.: OPTIMOL: automatic online picture collection via incremental model learning. Int. J. Comput. Vis. **88**, 147–168 (2009)
30. Schroff, F., Criminisi, A., Zisserman, A.: Harvesting image databases from the web. In: Proceedings of 11th IEEE International Conference Computer Vision (2007)
31. Felzenszwalb, P.F., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
32. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
33. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: describing images using 1 million captioned photographs. In: Advances in Neural Information Processing Systems (2011)
34. Kuznetsova, P., Ordonez, V., Berg, A.C., Berg, T.L., Choi, Y.: Collective generation of natural image descriptions. In: Annual Meeting of the Association for Computational Linguistics (2012)
35. Mason, R., Charniak, E.: Nonparametric method for data-driven image captioning. In: Annual Meeting of the Association for Computational Linguistics (2014)
36. Patterson, G., Xu, C., Su, H., Hays, J.: The SUN attribute database: beyond categories for deeper scene understanding. Int. J. Comput. Vis. 108 (1–2), 59–81 (2014)
37. Yagcioglu, S., Erdem, E., Erdem, A., Cakici, R.: A distributed representation based query expansion approach for image captioning. In: Annual Meeting of the Association for Computational Linguistics (2015)
38. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M.: Language models for image captioning: the quirks and what works. In: Annual Meeting of the Association for Computational Linguistics (2015)
39. Grangier, D., Bengio, S.: A discriminative kernel-based approach to rank images from text queries. PAMI **30**, 1371–1384 (2008)

40. Hodosh, M., Hockenmaier, J.: Sentence-based image description with scalable, explicit models. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (2013)
41. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.: Grounded compositional semantics for finding and describing images with sentences. Trans. Assoc. Comput. Linguist. **2**, 207–218 (2014)
42. Karpathy, A., Joulin, A., Fei-Fei, L.: Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in Neural Information Processing Systems (2014)
43. Sun, C., Gan, C., Nevatia, R.: Automatic concept discovery from parallel text and visual corpora. In: International Conference on Computer Vision (2015)
44. Pinheiro, P., Lebret, R., Collobert, R.: Simple image description generator via a linear phrase-based model. In: International Conference on Learning Representations Workshop (2015)
45. Ushiku, Y., Yamaguchi, M., Mukuta, Y., Harada, T.: Common subspace for model and similarity: phrase learning for caption generation from images. In: International Conference on Computer Vision (2015)
46. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. In: Advances in Neural Information Processing Systems Deep Learning Workshop (2015)
47. Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
48. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
49. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning (2015)
50. Lebret, R., Pinheiro, P.O., Collobert, R.: Phrase-based image captioning. In: International Conference on Machine Learning (2015)
51. Gupta, A., Kembhavi, A., Davis, L.: Observing human-object interactions: using spatial and functional compatibility for recognition. In: PAMI (2009)
52. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
53. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition (2006)
54. Prest, A., Schmid, C., Ferrari, V.: Weakly supervised learning of interactions between humans and objects. IEEE Trans. Pattern Anal. Mach. Intell. **34**(3), 601–614 (2012)
55. Verma, Y., Jawahar, C.V.: Im2Text and Text2Im: associating images and texts for cross-modal retrieval. In: British Machine Vision Conference (2014)
56. Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., Plank, B.: Automatic description generation from images: a survey of models, datasets, and evaluation measures. J. Artif. Intell. Res. **55**, 409–442 (2016)
57. Lenat, D.B.: Cyc: a large-scale investment in knowledge infrastructure. Commun. ACM **38**(11), 33–38 (1995P

Age-Type Identification and Recognition of Historical Kannada Handwritten Document Images Using HOG Feature Descriptors

Parashuram Bannigidad and Chandrashekar Gudada

Abstract Most of the historical Kannada handwritten documents are preserved in the manuscript preservation centre and archaeological departments. The historical Kannada handwritten documents are generally degraded in nature, due to this degradation, the documents are impossible to read and understand the contents. Hence, it is very much essential to restore by digitizing the historical Kannada handwritten documents and also recognize the originality of the dynasty to which it belongs. The main objective of the research work is to reconstruct, digitize and recognize the historical Kannada handwritten document images by applying image enhancement techniques and obtain the HOG feature descriptors using K-nearest neighbour (K-NN) and SVM classifiers. In this paper, we have considered historical Kannada handwritten document images of different dynasties based on their age-type; Vijayanagara dynasty (1460 AD), Mysore Wadiyar dynasty (1936 AD), Vijayanagara dynasty (1400 AD) and Hoysala dynasty (1340 AD) for experimentation. The average classification accuracy for different dynasties: in case of K-NN classifier is 92.3% and SVM classifier is 96.7%, It is observed that the SVM classifier has got a good classification performance comparatively K-NN classifier for Historical Kannada handwritten document images. The experimental outcomes are tested with manual results and other methods in the literature, which show the thoroughness of the proposed technique.

Keywords Restoration · Segmentation · Kannada · K-NN · SVM · Recognition HOG · Handwritten script · Historical documents

P. Bannigidad · C. Gudada (✉)
Rani Channamma University, Belagavi 591156, Karnataka, India
e-mail: chandrugudada@gmail.com

P. Bannigidad
e-mail: parashurambannigidad@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_101

1 Introduction

Kannada is the most prominently spoken and is the official language in the state of Karnataka in India, Most of the Indian dynasties were ruled in this area, and Kannada scripts are the oldest among 20 languages of the Dravidian family [1]. Kannada script is advanced over a time of 1500 years and there a lot of variations in the scripts used by different dynasties. The development of Kannada character has experienced numerous evolutions and changes amidst its use during the time of different dynasties [2]. The Kadamba script was the first dynastic rule of Karnataka in the fifth Century AD. The Adi Ganga script was recognized during Centuries between fourth AD and sixth AD. It has a resemblance to the Kadamba script. The script used by the Badami Chalukya is called Badami Chalukya script and it can be seen in the records of sixth—seventh Century AD. The Rastrakuta script recognized during eighteenth Century AD. The Kalyana Chalukya script was recognized between tenth and twelfth Century AD. The Vijayanagara script was used between fourteenth and sixteenth Century AD. The Mysore Wadiyar is the last dynasty, which ruled in 1eighteenthnineteenth Century AD, in this period, the Vijayanagara scripts were also evolved and it is almost similar to the Aadhunika Kannada script, which is also similar in nature to the present day Kannada handwritten script [3].

The restoration and recognition of degraded historical Kannada handwritten document images involve restoration, digitization, preservation, knowledge extraction, etc. In this paper, the degraded historical Kannada handwritten document images of Vijayanagara dynasty (1460 AD), Mysore Wadiyar dynasty (1936 AD), Vijayanagara dynasty (1400 AD) and Hoysala dynasty (1340 AD) written on a paper (Hasatprati) are considered, which are shown in Fig. 1. Many Government organizations; National Mission for Manuscript, New Delhi, and Archaeological Survey of India, manuscript preservation centres, universities, etc., are working on digitization process and preservation of degraded historical Kannada handwritten documents.

Kannada language inscriptions are categorized based on their style of writing as Pre-old Kannada (Poorva Halagannada), Old-Kannada (Halagannada), Nadugannada and Aadhunika Kannada. The evolution of sample historical Kannada handwritten scripts of different dynasties is shown in Fig. 2.

Very few researchers have contributed to this area in the literature. The recognition of holistic word from handwritten historical documents has been proposed by Lavrenko et al. [4]. The recognition of handwritten text form historical documents in the tranScriptorium project was investigated by Sanchez et al. [5]. A character recognition using geometry-based features was presented by Dileep et al. [6]. Romero et al. [7] proposed a handwritten text recognition for historical document. The recognition of words in historical, classical Mongolian document was investigated by Gao et al. [8]. The classification of ancient Kannada scripts using SVM and K-NN was carried out by Soumya et al. [9]. Prediction of the era of historical scripts using Curvelet transform-based approach was done by Gangamma et al. [10]. Predication of the era of a historical script using SVM classifier was carried out by Soumya et al. [11]. Gabor-and Zonal-based feature approach for recognition of historical documents

Fig. 1 Sample historical Kannada handwritten documents from different dynasties **a** Vijayanagara dynasty (1460 AD) **b** Mysore Wadiyar dynasty (1936 AD) **c** Vijayanagara dynasty (1400 AD) **d** Hoysala dynasty (1340 AD)

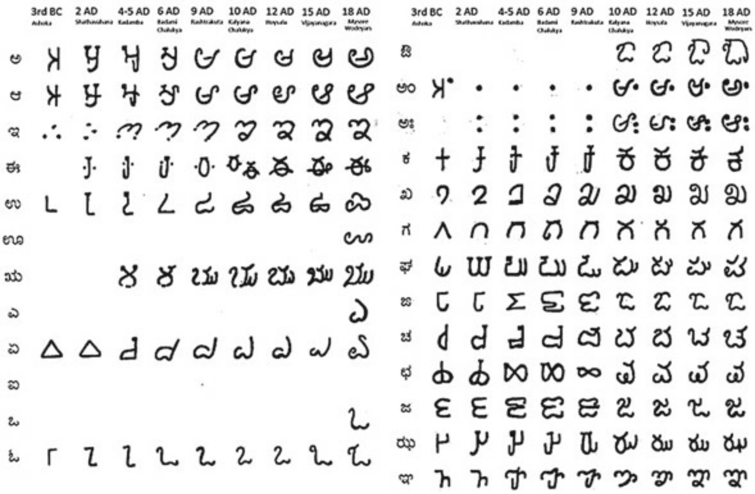


Fig. 2 The evolution of sample historical Kannada handwritten scripts of different dynasties

was carried out by Soumya et al. [12]. Stone-inscripted Kannada character matching using SIFTS has been investigated by Mohana et al. [13]. The offline handwriting text recognition was proposed by Verma et al. [14]. Parashuram and Chandrashekar [15, 16] have proposed an image enhancement method for degraded historical Kannada handwritten document images and ensure the quality of historic Kannada handwritten document images based on the performance evaluation approaches, i.e. Precision, Recall, F-Measure, MSE and PSNR.

2 Proposed Method

The goal of the present investigation is to reconstruct, digitize and recognize the historical Kannada handwritten document images by applying image enhancement techniques and obtain the HOG feature descriptors using K-nearest neighbour (K-NN) and SVM classifiers. The purpose of the recognition is to identify the script of the dynasties, whether it belongs to Hoysala dynasty or Vijayanagara dynasty or Mysore Wadiyar dynasty.

2.1 Data Collection

The standard datasets of historical Kannada handwritten documents are rarely found in the literature. The historical Kannada handwritten documents are collected individually by visiting many institutions and universities, like; Department of P. G. Studies and Research in Kannada, Gulbarga University, Kalaburgi and Dept. of Hasataprati, Kannada University, Hampi. These Kannada handwritten documents are captured through Canon 1300D, 18 megapixels DSLR Camera at 5184×3456 resolutions in the JPEG format and store them as soft copies (own datasets) which contains a total of 1200 images of historical handwritten Kannada documents.

2.2 Preprocessing

It is more common that, the historical documents contain smear, uneven background illumination, and spot due to age or marks resulting from the ink that goes through the inscription (paper), generally called bleed-through. Apart from this, the style of the writer varies with inscription to inscription, which leads to the confusion and complexity to recognize the historical document. The preprocessing steps which include image enhancement and restoration will play an important role, not only to enhance the quality of the image, but also removes the unwanted objects, debris, noise, etc.

In the earlier papers, we have proposed a new image enhancement technique for degraded historical Kannada handwritten document images and are described in [15, 16]. The detailed steps of the algorithm are described in Fig. 3. The detailed approach of the proposed method is discussed in the form of the algorithm as given below.

Algorithm: Age identification and classification of historical Kannada handwritten document images.

Step 1. Input the camera-captured colour image.
Step 2. Apply image enhancement techniques; namely the combination of Local Otsu and Global Otsu to restore and binarize the given input image.
Step 3. Apply text block-wise segmentation method on Step 2 of size 512×512 and store them individually.

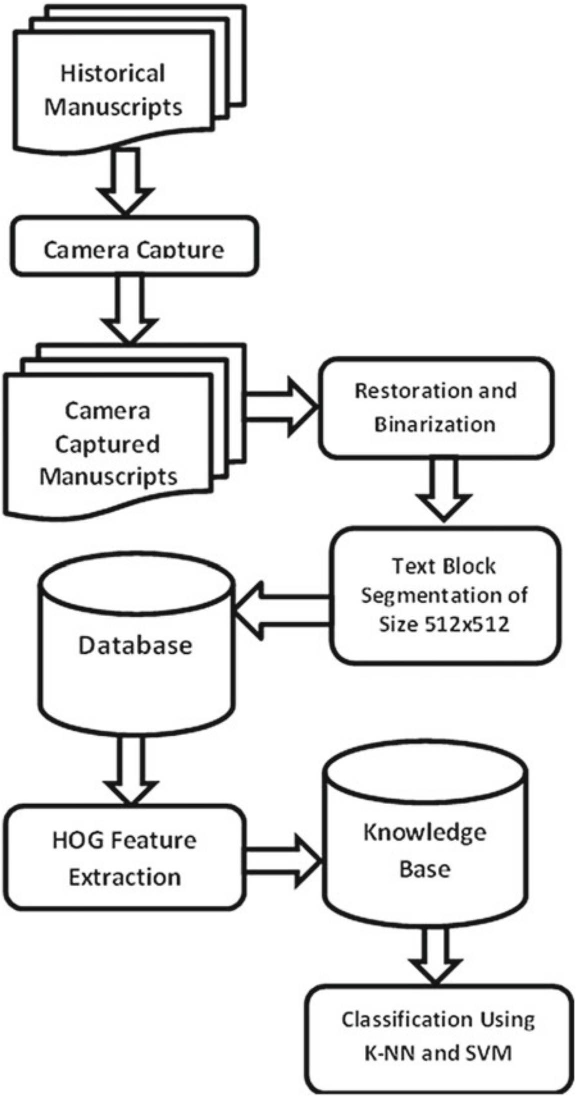


Fig. 3 The detailed approach of the proposed algorithm

Step 4. Repeat the Steps from 1–3 for all the Kannada handwritten document images of different age-types: namely Hoysala, Vijayanagara and Mysore dynasties

Step 5. Extract HOG feature descriptors from Step 4 and store them as a knowledge base

Step 6. Apply classification techniques; i.e. K-nearest neighbour classifier [17] and SVM [18] classifier to classify and recognize the historical Kannada handwritten document images, whether they belong to the Hoysala dynasty or Vijayanagara dynasty or Mysore dynasty.

2.3 Feature Selection and Extraction

The selection of features for document image is an important task, since the Kannada handwritten document image contains different shapes and styles. In this paper, we have used HOG feature descriptors [19]; and these are extracted based on their scale invariant and rotation invariant for recognition of the Kannada handwritten documents.

3 Experimental Results and Discussion

For the of purpose experimentation, we have considered 1200 image datasets from different dynasties, namely; Vijayanagara, Hoysala, and Mysore Wadiyar are described in Sect. 2.1.

The implementation is done on Intel Core i5, @ 2.40 GHz system using MATLAB R2015b. The original Historical Kannada handwritten camera captured document image of Vijayanagara dynasty (1460 AD) (Fig. 4a), converts the camera captured colour RGB image into greyscale image and considers only green channel and the combination of local otsu and global otsu is applied to reconstruct the individual character and output the enhanced and binarized image (Fig. 4b), apply the block segmentation method on Fig. 4b to obtain individual text block of size 512×512 (Fig. 4c). Extract the HOG feature descriptors on Fig. 4c for all the images and store them as a knowledge base. Finally, apply classification techniques; i.e. K-NN classifier and SVM classifier for classification and recognition. The average classification accuracy of the proposed method with K-NN classifier and SVM classifier is given in Table 1. The comparative performance of the proposed method in the literature is given in Table 2.

The classification accuracy for different dynasties: the K-NN classifier has got 92.3% and SVM classifier has got 96.7%. Based on the experimentation [23, 24], it is observed that the SVM classifier has got a good classification performance comparatively K-NN classifier for historical Kannada handwritten document images. The results of the confusion matrix of K-NN classifier and its performance accuracy towards historical Kannada handwritten document images is given in Table 3. The result of the confusion matrix of SVM classifier which, is shown in Table 4 indicates better recognition rates towards historical Kannada handwritten document images.

Table 1 The classification accuracy of the proposed method with K-NN and SVM classifier

Age-types (Dynasties)	Classification accuracy of the proposed method			
	K-NN classifier		SVM classifier	
	Recognition rate (%)	Error rate (%)	Recognition rate (%)	Error rate (%)
Vijayanagara (1460 AD)	98	2	98	2
Mysore Wadiyar (1936 AD)	86	14	96	4
Vijayanagara (1400 AD)	88	12	96	4
Hoysala (1340 AD)	98	2	97	3
Average accuracy	92.3		96.7	

Table 2 The comparison performance of the proposed method in the literature

Author	Method	Dataset size	Accuracy (%)
Karthik et al. [20]	Adaptive window sizing and histogram of oriented gradient	200	94
Alaci et al. [21]	Potential piece-wise separation line technique	204	94.98
Belagali et al. [22]	Zoning-based invariant moment feature	980	94.69
Proposed method	Histogram-oriented gradient feature descriptors	1200	96.70

Table 3 Confusion matrix for K-NN classifier

Age-types (Dynasties)	Vijayanagara (1460 AD)	Mysore Wadiyar (1936 AD)	Vijayanagara (1400 AD)	Hoysala (1340 AD)	Unknown	Total
Vijayanagara (1460 AD)	294	1	–	5	–	300
Mysore Wadiyar (1936 AD)	4	257	1	38	–	300
Vijayanagara (1400 AD)	9	–	264	27	–	300
Hoysala (1340 AD)	5	–	2	293	–	300

Fig. 4 Sample images of the proposed algorithm **a** Original camera captured historical Kannada handwritten document image of Vijayanagara dynasty (1460 AD) **b** Binarized image after restoration **c** Text block segmentation of size 512 × 512 on (**b**)

Table 4 Confusion matrix for SVM classifier

Age-types (Dynasties)	Vijayanagara (1460 AD)	Mysore Wadiyar (1936 AD)	Vijayanagara (1400 AD)	Hoysala (1340 AD)	Unknown	Total
Vijayanagara (1460 AD)	294	4	–	2	–	300
Mysore Wadiyar (1936 AD)	2	288	1	9	–	300
Vijayanagara (1400 AD)	7	–	287	6	–	300
Hoysala (1340 AD)	3	1	5	291	–	300

4 Conclusion

In this paper, we have proposed an algorithm to restore and recognize handwritten documents from the historical Kannada handwritten documents by extracting HOG feature descriptors and accuracy is measured by using K-NN and SVM classifiers. For the purpose of experimentation, we have considered historical Kannada handwritten document images of different dynasties based on their age-type; i.e. Vijayanagara dynasty (1460 AD), Mysore Wadiyar dynasty (1936 AD), Vijayanagara dynasty (1400 AD) and Hoysala dynasty (1340 AD). The average classification accuracy for different dynasties: the K-NN classifier has got 92.3% and SVM classifier has got 96.7%, Based on the experimentation, it is observed that the SVM classifier has got a good classification performance comparatively K-NN classifier for historical Kannada handwritten document images. The experimental outcomes are tested with manual results and other methods in the literature, which shows the thoroughness of the proposed technique.

Acknowledgements The authors are indebted to The Chairman, Department of P. G. Studies and Research in Kannada, Gulbarga University, Kalaburgi and Dept. of Hasataprati, Kannada University, Hampi for providing the historical Kannada handwritten documents and perception of manual outcomes.

References

1. Manjunath, M.G., Devarajaswamy, G.K., Vikasa, K.L.: Jagadhguru Sri Madhvacharya Trust, Sri Raghavendra Swami Matta, Mantralaya
2. Narasimha Murthy, A.V.: Kannada Lipiya Ugama Mattu Vikasa, Kannada Adhyayana Samsthe, Mysore University, Mysore (1968)
3. Reddy, D.: Lipiya Huttu Mattu Belavanige—Origin and Evolution of Script, Kannada Pustaka Pradhikara (Kannada Book Authority), Bangalore
4. Lavrenko, V., Rath, T.M., Manmath, R.: Holistic word recognition for handwritten documents. In: Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04), pp. 278–287 (2004)
5. Sanchez, J.A., Bosch, V., Romero, V., Depuydt, K., de Does, J.: A Handwritten text recognition for historical documents in the tran Scriptorium Project. In: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, DATeCH '14, pp. 111–117 (2014)
6. Dileep, D.: A feature extraction technique based on character geometry for character recognition. Comput. Res. Repos. J. **1202**(3884), 1–4 (2012)
7. Romero, V., Serrano, N., Toselli, A.H., Sanchez, J.A., Vidal, E.: Handwritten text recognition for historical documents. In: Proceedings of Langauage Technologies for Digital heritage and Cultural Heritage Workshop, pp. 90–96
8. Gao, G., Su, X., Wei, H., Gong, Y.: Classical Mongolian words recognition in historical document. IEEE ICDAR **1520–5363**(11), 692–697 (2011)
9. Soumya, A., Hemantha Kumar, G.: Performance analysis of random forests with SVM and KNN in classification of ancient Kannada scripts. Int. J. Comput. Technol. **13**(9), pp. 4907–4921 (2014)

10. Gangamma, B., Murthy, K.S., Punitha, P.: Curvelet transform based approach for prediction of epigraphical scripts era. In: IEEE International Conference on Computational Intelligence and Computing Research, pp. 1–6 (2012) 978-4673-1344/12
11. Soumya, A., Hemantha Kumar, G.H.: SVM classifier for the predication of era of an epigraphical script. IJP2P **2**(2), 12–22 (2011)
12. Soumya, A., Hemantha Kumar, G.: Recognition of historical records using Gabor and zonal features. SIPIJ **6**(4) 57–69 (2015)
13. Mohana, H.S., Navya, K., Srikanth, P.C., Shivakumar, G.: Stone inscripted Kannada Character matching Using SIFT. In: Proceedings of the IRF International Conference, pp. 126–131 (2014)
14. Verma, B., Blumenstein, M., Kulkarni, S.: Recent achievements in offline handwriting recognition system. In: International Conference on Computational Intelligence and Multimedia Applications (1998)
15. Bannigidad, P., Gudada, C.: Restoration of degraded historical Kannada handwritten document images using image enhancement techniques. In: International Conference on Soft Computing and Pattern Recognition (SoCPaR 2016), pp. 498–508. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-60618-7_49
16. Bannigidad, P., Gudada, C.: Restoration of degraded Kannada handwritten paper inscriptions (Hastaprati) using image enhancement techniques. In: IEEE International Conference on Computer Communication and Informatics (ICCCI-2017), pp. 1–6 (2017) https://doi.org/10.1109/iccci.2017.8117697
17. https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/
18. http://www.statsoft.com/Textbook/Support-Vector-Machines#Classification SVM
19. https://en.wikipedia.org/wiki/Histogram_of_oriented_gradients
20. Karthik, S., Srikanta Murthy, K.: Segmentation and recognition of handwritten kannada text using relevance feedback and histogram of oriented gradients—a novel approach. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **7**(1), 472–476 (2016)
21. Alaei, A., Nagabhushan, P., Pal, U.: A benchmark Kannada handwritten document dataset and its segmentation. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 141–145 (2011)
22. Belagali, Netravati, Angadi, Shanmukhappa A.: OCR for handwritten Kannada language script. Int. J. Recent Trends Eng. Res. (IJRTER) **02**(08), 190–197 (2016)
23. Sridharamurthy, S.K., Sudarshana Reddy, H.R.: PCA based feature for handwritten Kannada characters recognition. In: International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT) (2015). https://doi.org/10.1109/erect.2015.7499053
24. Manjunath Aradhya, V.N., Hemantha Kumar, G., Noushath, S.: Multilingual OCR system for South Indian scripts and English documents: an approach based on Fourier transform and principal component analysis. In: Engineering Applications of Artificial Intelligence Elsevier, vol. 21, no. 4, pp. 658–668 (2008)

Image Inpainting for Hemorrhage Detection in Mass Screening of Diabetic Retinopathy

Anupama Awati, H. Chinmayee Rao and M. R. Patil

Abstract Diabetic retinopathy (DR) is one of the main causes of vision loss. The severity of DR can be analyzed using human retinal images (fundus image). Affected fundus image consists of hemorrhages, microaneurysms, and exudates along with blood vessels. In order to accurately detect the level of severity of the disease, the blood vessels are inpainted using fast marching method (FMM). The technique implemented in this paper involves image enhancement using green channel image and histogram equalization followed by mask generation and inpainting. The severity of the disease can be categorized accurately by inpainting the blood vessels using FMM. The proposed technique is tested using standard test databases HRF and DRIVE. The algorithm can be effectively used for mass screening of DR. This technique is a fundamental step in designing computer-aided diagnosis system for ophthalmic disorders.

Keywords Retinopathy · Image inpainting · Fundus image · Mass screening

1 Introduction

Human retinal images form the basis for diagnosis of various pathologies. In ophthalmology, the interior surface of the eye containing retina, optic disc, blood vessel, fovea, and macula [1] is called as fundus of eye which can be clearly observed and distinguished from Fig. 1a. Various pathologies are also present in eye fundus. One among them is diabetic retinopathy. Usually, early DR symptoms are microaneurysms, hemorrhages, and exudates [2]. An example of affected fundus image with these symptoms can be seen in Fig. 1b. It is a retinal disorder which leads to vision

A. Awati (✉) · H. C. Rao
Department of Electronics and Communication, KLS Gogte Institute of Technology, Udyambag, Belagavi, India
e-mail: asawati@git.edu

M. R. Patil
Department of Electronics and Communication, JAGMIT, Jamakhandi, India

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_102

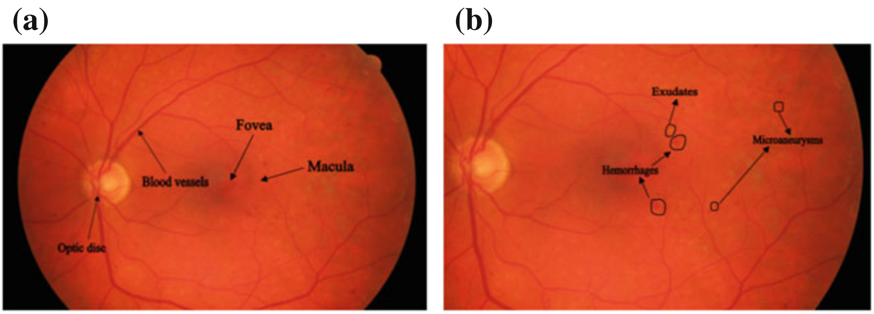


Fig. 1 **a** Parts of fundus images **b** Hemorrhage, exudates and microaneurysms

impairment. It causes the blood vessels in the retina to leak or become blocked and damage our eyesight. Basically, there are two stages in diabetic retinopathy. The first stage is known as Non-Proliferative Diabetic Retinopathy (NPDR). And the second stage is Proliferative Diabetic Retinopathy (PDR). In the initial stages, there are no significant symptoms. Hence early detection of this disease is difficult. In the second stage, abnormal new blood vessels are formed which later burst and bleed causing blur vision. These can be detected in the fundus images as exudates and hemorrhages as shown in Fig. 1b. Exudates are the sediments produced by the leaked fluid compound of lipid byproducts and appear as white/yellow soft structures in color fundus images [3]. Hemorrhages occur due to the leaking of false blood vessels formed in retina [2]. Hemorrhages are one of the most significant signs during diabetic retinopathy detection. Hence, developing an automatic detection of hemorrhages will aid in the diagnosis of DR.

In diabetic retinopathy screening, inpainting blood vessels from the fundus images play an important role for further classification of the DR stage. For hemorrhage detection, the blood vessels, optic disc, and fovea in the retinal images are included in background. So except for this, all the other objects can be considered as foreground [2]. Most of the literatures that have published detect hemorrhages without background blood vessel removal. Figure 1b clearly shows that how the hemorrhages appear similar to the color of blood vessels and darker than background. Inpainting blood vessel improves the efficiency of detection of these hemorrhages. In this paper, we concentrate on removing of blood vessels from the fundus images which help in easy detection of hemorrhage.

Inpainting is a process of reconstruction of lost or deteriorated parts of images or removal of unwanted segments of image [4]. Researchers have developed several algorithms to obtain a good reconstructed image. These algorithms can be broadly classified as diffusion-based method and exemplar-based method [4]. Diffusion-based methods involve complex implementation process and the inpainting process is slow. These methods are useful in restoring straight lines and curves or filling small gaps. In exemplar-based methods, the unknown regions of image are filled by selecting the best matching patch from known regions of image [4]. These methods are largely used due to its fast computational speed but give good results for filling large regions. Apart from these, there are convolution based algorithms which use weighted kernels. These algorithms are fast compared to other two methods. Inpainting is performed by convolving the fundus image with the kernel. For every pixel, kernel weights are calculated using its neighbors in both space and range domain to perform bilateral filtering.

2 Literature Review

Most of the literatures that have published hemorrhage and exudates detection have carried out experiments without removing blood vessels in fundus images [2, 3]. Paper [2] concentrates on improving the detection of hemorrhages and making the detection system more robust by using background estimation and Mahalanobis distance. Their method has been able to reduce the false negative results that arise due to inaccurate vessel structure by vessel exclusion method. To avoid the errors due to nonuniform illumination in image M. Foracchia's correction technique is used along with normalization. They have applied vessel exclusion technique using shape analysis. Since the background estimation may include vessel and noise pixels, these have been removed by excluding candidate lesions less than 10 pixels. The blood vessel in retinal images causes errors in hemorrhage detection. This can be avoided by inpainting blood vessels as in [4, 5]. Some researchers go for [5] inpainting blood vessels and other unwanted segments or candidates from the retinal image using methods like morphological inpainting technique.

In [4], blood vessel inpainting is the major focus and achieved using sparse representations and dictionary learning techniques. The results were evaluated using two methods (1) by traditional recovery error (2) by adding false vessels in the fundus images. Paper [3] focuses on extraction of hemorrhages and exudates from DR affected fundus images by using techniques like Circular Hough Transform (CHT), and Gabor filtering. Their algorithm separates hard exudates and soft exudates and hemorrhages separately. In [6], blood vessel extraction is done using morphological closing operation. After the blood vessel extraction, candidate detection is done using Ostu's Threshold on the green channel image. For the detection lesions and classification, random forest approach is used.

Paper [7] proposes a novel method for detection of exudates by superpixel multi-feature classification algorithm approach. It is done by dividing entire image into series of superpixels considered as candidates. This approach also distinguishes true exudates from the spurious candidates. In paper [8], blood vessels are inpainted for detection and segmentation of optic disc in digital fundus images. They have proposed two segmentation techniques one for retinal vessel network segmentation and other for optic disc segmentation. Median filtering approach has been implemented for inpainting blood vessels efficient optic disc segmentation obtaining average coefficient of correlation 0.918. Paper [9] proposes median diffusion-based image inpainting technique. The technique is verified for both homogenous and heterogeneous backgrounds and performance is analyzed based on PSNR values (peak signal to noise ratio). Our work is mainly inspired by [5] and uses inpainting techniques similar to [10].

3 Methodology

In this paper, we propose algorithm for retinal blood vessel removal using image inpainting for improved detection and classification of hemorrhages in the fundus images. The algorithm for the proposed method is as follows:

(1) Fundus image acquisition.
(2) Image preprocessing (Green channel image).
(3) Contrast enhancement.
(4) Mask generation.
(5) Image inpainting.

The preprocessing stage involves image enhancement and segmentation for mask image generation. Image preprocessing and mask generation: For diabetic retinopathy screening should obtain fundus images with good contrast so that the blood vessels are distinctly visible compared to background. So the first step is to obtain a green channel image [3] of the original fundus image. This step is essential since the green channel image gives the best vessel background contrast. Further, the contrast is improved using contrast limited adaptive histogram equalization (CLAHE) [11] method of the green channel image. CLAHE method is an image contrast enhancement method which differs from the normal histogram equalization by performing several histograms of pertaining to different sections of image and avoid over-amplification of noise by limiting the contrast enhancement of Adaptive histogram equalization. Mask images help us identify the location in which the image has to be inpainted. Mask generation is achieved by applying alternate sequential filtering by closing opening as suggested in paper [4]. Removing very small contours through area parameter and noise removal is implemented followed by removing blobs of unwanted bigger chunks taking in consideration they are not straight lines like blood vessels and also in an interval of area. Figure 3 shows the mask generated using the above technique.

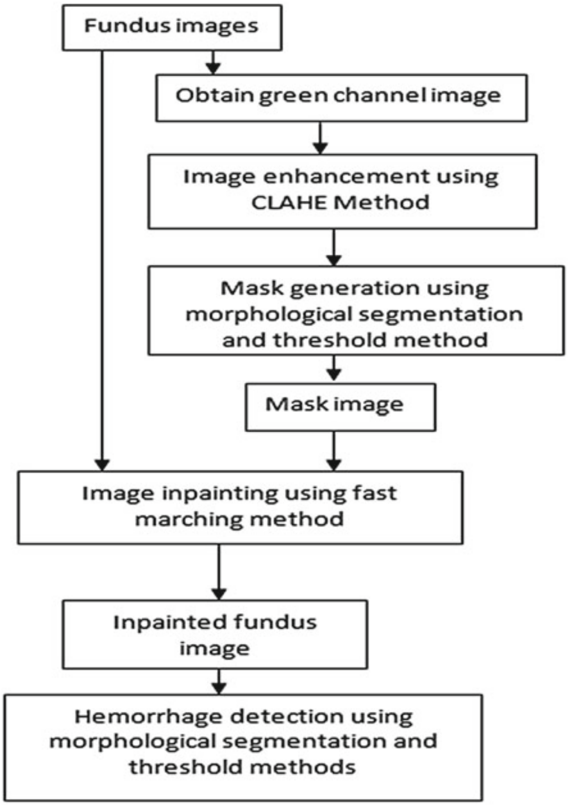
Blood vessel inpainting: For inpainting blood vessels, we will be using Telea method which based on Bertalmios equation [12] and Fast Marching Method (FMM).

Consider a region Ω to be inpainted and let $\partial\ \Omega$ be its boundary.

The region to be inpainted is specified by the mask image generated in the previous step. Consider a small neighborhood region $B_\varepsilon(p)$ of size ε around a point p on the boundary $\partial\Omega$. For small ε, $I_q(p)$ of the image at point p, given the image I(q) and gradient ∇(q) of point q (Fig. 2)

$$I_q(p) = I(q) + \nabla I(q)(p - q) \tag{1}$$

Fig. 2 Flowchart for the proposed method



This equation gives pixel value at p only due to one point in neighborhood. But it will give better results if we consider all points defined by region $B_{\varepsilon}(p)$. So we have to inpaint point p as a function of all points q in the neighborhood region by summing the estimates of all the points in neighborhood, weighted by a normalized function w(p, q):

$$I(p) = \frac{\sum_{q \in B\varepsilon(p)} w(p, q)[I(q) + \nabla I(q)(p - q)]}{\sum_{q \in B\varepsilon(p)} w(p, q)} \tag{2}$$

The weighting function $w(p, q)$ is calculated based on the direction vector distance from the boundary and the level set. Since we have to inpaint whole Ω, we need to iteratively compute all the pixels on the boundary ∂ Ω starting from initial pixel at p. To advance the boundary inside Ω, we make use of FMM technique. FMM algorithm solves the Eikonal equation $|\nabla T| = 1$, with $T = 0$ on the boundary $\partial\Omega$. Solution of this equation gives the distance map of Ω pixels with respect to boundary. Based on this distance maps, the level sets are determined using which we can easily propagate the image information from the neighborhood to the center of the

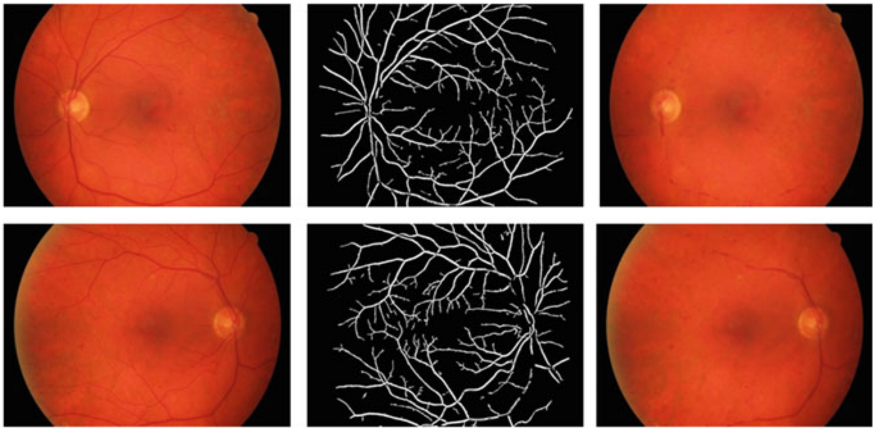


Fig. 3 Original fundus images. Mask images inpainted images

inpainting region. This technique ensures that we always proceed in increasing order of their distance to boundary. The main advantage of this method is that the level set need not be connected, and can break and merge in next advances.

The last step is to extract the hemorrhages from the inpainted retinal images. We have set the threshold in accordance with the candidate to be extracted. Here, the candidate is hemorrhage. Further morphological characteristics are used to segment hemorrhages.

4 Result

We have implemented the proposed method on High Resolution Fundus (HRF) Image and STARE Databases in Python2.7. The sample images are as shown in Fig. 3 where the first column contains original image, second column contains corresponding mask image, and third column contains corresponding inpainted images. From each of these inpainted images, we have calculated the number of pixels p in the image affected with hemorrhages. From affected pixel count we find the ratio of p by total number of pixels, which gives an estimate of severity of the disease. The results are compared with two other inpainting methods. Method I is neighborhood-based inpainting and method II is exemplar-based inpainted. Blood vessels inpainted using these methods have been used for comparison. The results are tabulated in Table 1.

Table 1 Estimation of approximate severity of Diabetic retinopathy

Test image (HRF database)	No of pixels affected with hemorrhage (p)			Ratio p/t, where t is the total number of pixels			Percentage severity		
	Method			Method			Method		
	I	II	Our method	I	II	Our method	I	II	Our method
1	14,998	13,056	12,145	0.00183	0.00159	0.00148	0.1832	0.1595	0.1484
2	30,467	29,012	24,996	0.00372	0.00354	0.00357	0.3721	0.3544	0.3575
3	4,41,727	4,36,949	4,11,945	0.05396	0.05338	0.05032	5.3965	5.3381	5.0327
4	4,35,890	4,25,985	3,73,279	0.05325	0.05325	0.04560	5.3252	5.3252	4.5603
5	3480	678	214	0.00042	0.00082	0.00002	0.0425	0.0082	0.0026

5 Conclusion

This paper presents an efficient method for mass screening for diabetic retinopathy on large scale by using image inpainting technique. In this technique, we are propagating from the boundary of inpainting area to the center using FMM technique. Inpainting blood vessels effectively improve the detection of hemorrhage. This type of screening method is fast and efficient in screening large amount of retinal images more accurately. It can be extended to easy classification of diabetic retinopathy stages and analyze the severity of the disease. We can extend this technique for exudates and microaneurysms also and identify the stage of diabetic retinopathy using Image classification methods.

References

1. Bugeau, A., Bertalmío, M., Caselles, V., Sapiro, G.: A comprehensive framework for image inpainting. IEEE Trans. Image Process. **19**(10) (2010)
2. Zhang, D., Li, X., Shang, X., Yi, Y., Wang, Y.: Stand, robust hemorrhage detection in diabetic retinopathy image. IEEE Trans. Image Process. (2011). 978-1-4577-0121-4/11/
3. Biran, A., Bidari, P.S., Raahemifar, K.: Automatic method for exudates and hemorrhages detection from fundus retinal images. World Acad. Sci. Eng. Technol. Int. Eng. **10**(9) (2016)
4. Colomera, A., Naranjoa, V., Engan, K., Skretting, K.: Assessment of sparse-based inpainting for retinal vessel removal. Signal Process. Image Commun. Sci. Direct (2017)
5. Zhang, X., Thibault, G., Decencière, E., Marcotegui, B., Laÿ, B., Danno, R., Cazuguel, G., Quellec, G., Lamard, M., Massin, P., Chabouis, A., Victor, Z., Erginay, A.: Exudate detection in color retinal images for mass screening of diabetic retinopathy. Med. Image Anal. (2014)
6. Kaur, N., Kaur, J., Acharyya, M., Kapoor, N., Chatterjee, S., Gupta, S.: A supervised approach for automated detection of hemorrhages in retinal fundus images. In: IEEE Trans. (2016)
7. Zhou, W., Wu, C., Yi, Y., Du, W.: Automatic detection of exudates in digital color fundus images using super pixel multi-feature classification. IEEE Trans. 2169–3536 (2017)
8. Pavani, T.C., Pardhasaradhi, M.: Efficient localization of blood vessel in painting based technique and segmentation of optic disc in digital fundus images. Int. J. Adv. Technol. Innov. Res. **08**(01), pp. 0111–0116, Jan 2016
9. Biradar, R.L., Kohir, V.V.: A novel image inpainting technique based on median diffusion. S¯adhan¯ a, vol. 38, Part 4, pp. 621–644. c Indian Academy of Sciences, Aug 2013
10. Telea, A.: An Image Inpainting Technique Based on the Fast Marching Method (2004)
11. Zuiderveld, K.: Contrast limited adaptive histogram equalization. In: Heckbert, P. (ed.) Graphics Gems IV. Academic Press (1994)
12. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Akeley, Kurt (ed.) Proceedings SIGGRAPH 2000, Computer Graphics Proceedings, Annual Conference Series, pp. 417–424. Addison-Wesley, Reading, MA (2000)
13. Budai, A., Bock, R., Maier, A., Hornegger, J., Mishelson, G.: Robust vessel segmentation in fundud images. Int. J. Biomed. Imaging **2013** (2013)
14. Staal, J.J., Abramoff, M.D., Niemeijer, M., Viergever, M.A., van Ginneken, B.: Ridge based vessel segmentation in color images of retina. IEEE Trans. Med. Imaging **23**, 501–509 (2004)
15. Sreejini, K.S., Govindan, V.K.: A review of computer aided detection of anatomical structures and lesions of DR from color retina images. Int. J. Images Graph. Signal Process. (2015)
16. De, C., Shih, F.Y.: Improved Image Inpainting using Maximum Value Edge Detector. IEEE (2012). 978-1-4673-1830-3/12/$31.00 ©2012

17. Noori, H., Saryazdi, S., Nezamabadi-pour, H.: A convolution based image inpainting. In: International Conference on Communication Engineering, University of Sistan and Baluchestan, pp. 22–24 (2010)
18. Bertalmio, M., et al.: Image inpainting. In: Proceeding of SIGGRAPH 2000 Computer Graphics Processings, pp. 417–424
19. Hadhoud, M.M., Moustafa, K.A., Shenoda, S.Z.: Digital images inpainting using modified convolution based method. Int. J. Signal Process. Image Process. Pattern Recogn. 1–10
20. Elango, P., Murugesan, K.: Digital image inpainting using cellular neural network. Int. J. Open Probl. Comput. Sci. Math. **2**(3), 439–449 (2009)

Performance Analysis and Implementation of DES Algorithm and RSA Algorithm with Image and Audio Steganography Techniques

Ankit Gambhir, Khushboo and Rajeev Arya

Abstract In today's era, data security is an important concern. It is most demanding issue nowadays. It is essential for people using online banking, e-shopping, reservations, etc. The two major techniques that are used for secure communication are Cryptography and Steganography. Cryptographic algorithms scramble the data, so that the intruder will not be able to retrieve it; however, steganography covers that data in some cover file, so that the presence of communication is hidden. There are some techniques that integrate cryptography and steganography to provide multi-layer security. This paper shows the comparison of such two techniques. These are "RSA cryptography with image steganography and with audio steganography" and "DES cryptography with image steganography and with audio steganography". Stimulated results have been presented using MATLAB software.

Keywords Cryptography · Steganography · Data security · Intruder · RSA DES · Image steganography · Audio steganography

1 Introduction

Internet users in India have increased enormously. In this digital era, people are glued to the internet. From money transfer to shopping, movie ticket booking to railway ticket booking, and food ordering to bill payments, all activities requires transfer of confidential data like ATM pin, OTP, user IDs, etc. [1]. Information security is crucial

A. Gambhir
Delhi Technical Campus, Noida, India
e-mail: er.ankit.gambhir@gmail.com

Khushboo
Northern India Engineering College, New Delhi, India
e-mail: kkhushboo_2008@yahoo.com

R. Arya (✉)
National Institute of Technology, Patna, India
e-mail: rajeev.arya.iit@gmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_103

nowadays. Cryptography and Steganography are the techniques used for information security [2]. Cryptography scrambles the data, however, steganography hides the presence of data [3]. So far, both the techniques are independently used to secure information. However, there are some research papers available that merges both the techniques to provide multi-layer security [4, 5]. RSA cryptography algorithm is integrated with image steganography and with audio steganography separately, so that more level of security is maintained and risk of the intruder is mitigated [6]. In this paper, combination of "DES cryptography algorithm with image steganography and with audio steganography" and combination of "RSA cryptography algorithm with image steganography and with audio steganography" are implemented and simulation results have been shown.

2 Related Work

2.1 RSA Algorithm

In 1977, Ron Rivest, Adi Shamir, and Len Adlemen developed an RSA algorithm. It is based on the fact that "two prime numbers can be easily multiplied by cannot be easily factorize" [7]. RSA is an asymmetric key cryptographic algorithm, i.e., keys used for encryption and decryption is different.

The procedure of RSA algorithm is:

- Select two large prime numbers X & Y (say) so that $X \neq Y$.
- Determine P, by multiplying X & Y.
- Determine L by formula $L = (X - 1) * (Y - 1)$.
- Select a public key K such that K is not the factor of L.
- Choose private key D such that $(D * K) \bmod M = 1$.
- To calculate ciphertext (C): $C = M^K \bmod P$.
- To determine plain text (M): $M = C^D \bmod P$.

2.2 DES Algorithm

DES is data encryption standard. For encryption and decryption, a similar key is used [7].

- 64-bit plain text is taken as input and 56-bit key and creates output 64-bit block.
- The plaintext goes through an initial permutation, (IP). It experiences an inverse final permutation at the closing stages (IP^{-1}).
- The plain text that passes through an IP to produce the permutated bit.
- Halves of permuted block are produced by IP—left plain text and right plain text.
- With its own keys, 16 rounds of encryption are done.

(a) (b)

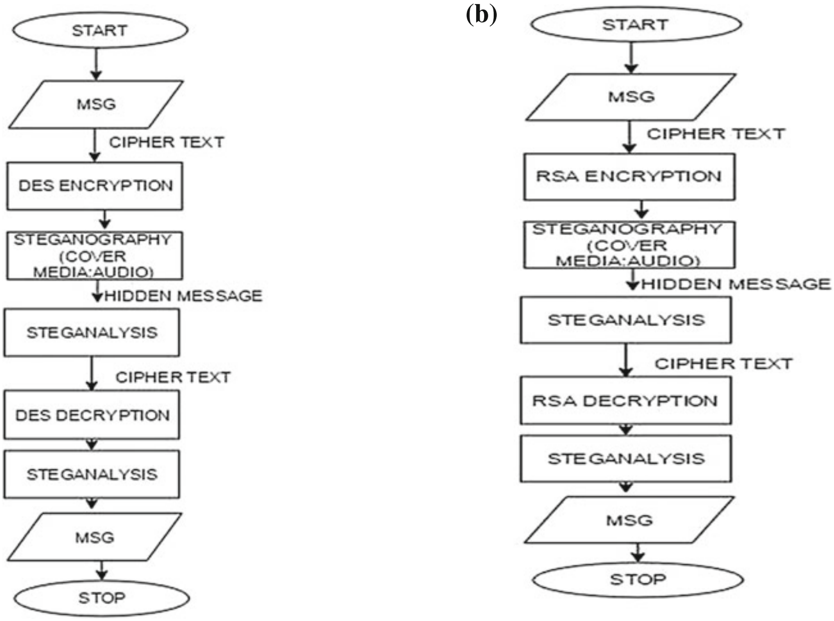


Fig. 1 **a** Flowchart Case 1, **b** Flowchart Case 2

- The output of 16 rounds of encryption consists of 64 bits of plain text and key.

64-bit ciphertext is produced at the last stage when the output undergoes inverse final permutation (IP^{-1}).

2.3 *Image Steganography and Audio Steganography*

In steganography, data is hidden in some cover file. Cover file could be text, image, audio, and video [3]. A very well-liked method is LSB, i.e., Least Significant Bit algorithm [8]. In this, LSB of cover files (image and audio) are replaced by bytes of data file [9–11].

3 Implementation

There are two cases. In Case 1, message has to be covered in an audio/image file is first converted in to ciphertext by using RSA cryptography algorithm (Fig. 1a).

In Case 2, message that has to be covered in an audio/image file is first converted in to ciphertext by using DES cryptography algorithm (Fig. 1b).

(a)

(b)

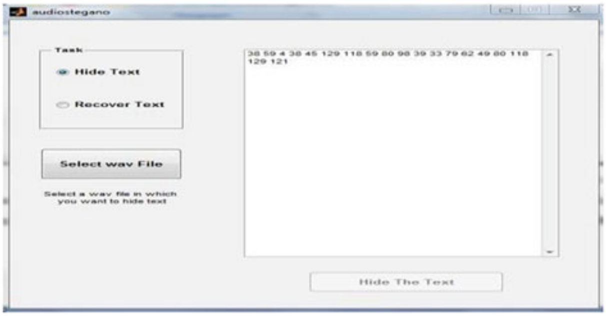


Fig. 2 **a** RSA encryption output hidden in image, **b** GUI for Audio Steganography

Case 1: RSA Algorithm with Image and Audio Steganography
Entered message is "**Hello**" and value of X and Y is taken as 11 and 13, respectively. The corresponding ciphertext evaluated by RSA algorithm is "**38 59 4 38 45 129 118 59 80 98 39 33 79 62 49 80 118 129 121**".

Outcome of RSA algorithm (ciphertext) is taken as input and hidden in image (Fig. 2).

Output of RSA algorithm (ciphertext) is taken as input and hidden in an audio.

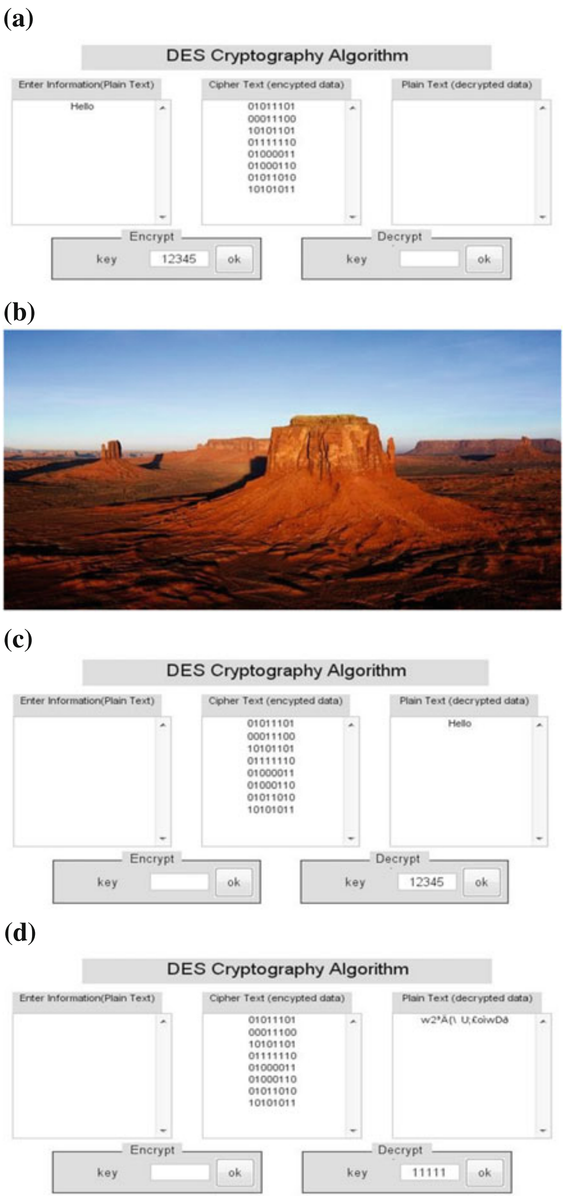
To recover ciphertext from audio, click recover text and select audio in which ciphertext is hidden.

Case 2: DES Algorithm with Image and Audio Steganography
Entered message is again "**Hello**" and selected key is "12345". Corresponding ciphertext is obtained (Fig. 3).

Output of DES algorithm (ciphertext) is hidden in image.

Ciphertext obtained by DES algorithm is entered to hide it in cover (audio) file. The ciphertext is saved in an audio "new254". To recover ciphertext from audio, recover text option is clicked and audio "new254" is selected in which ciphertext is hidden.

Fig. 3 **a** GUI for DES Cryptography Algorithm, **b** DES encryption output hidden in image, **c** GUI for DES Cryptography Algorithm, **d** GUI for DES Cryptography Algorithm



The recovered ciphertext is entered in GUI of DES Algorithm and same key is used to decrypt as DES is symmetric key algorithm. Same message, i.e., "**Hello**" is retrieved.

If key does not match, then message will not retrieved.

4 Result Analysis

Result analysis of all audios; without data, with RSA ciphertext and with DES ciphertext.

It can be observed from the Fig. 4a–c that all the three waveforms are indistinguishable. Characteristics of audio signal after being embeded with data do not vary. Histogram of all the images; without data, with RSA ciphertext and with DES ciphertext are presented in the Fig. 4d.

With the increase in the use of social media by the users, nowadays, cloud-based multimedia data storage and processing are becoming very popular. Hence, the storage and data security of such a huge data are a prime concern for the cloud service providers [12–15]. The proposed methodology can be used as a security tool in aforementioned scenarios.

5 Conclusion

In this paper, "RSA cryptography algorithm with image steganography and with audio steganography" and "DES cryptography algorithm with image steganography and with audio steganography" have been studied and implemented. Stimulated results have been shown. All three waveforms are identical. These techniques are better than individual techniques. The risk of unauthorized access is alleviated up to a certain extent by using these techniques. These techniques could be used in banks or agencies where highly confidential data is transferred. There is a broad scope of future work, for instance, a novel algorithm could be design by amending RSA, DES, or AES cryptographic algorithm,and amendment could also be made in LSB embedding techniques to make it robust.

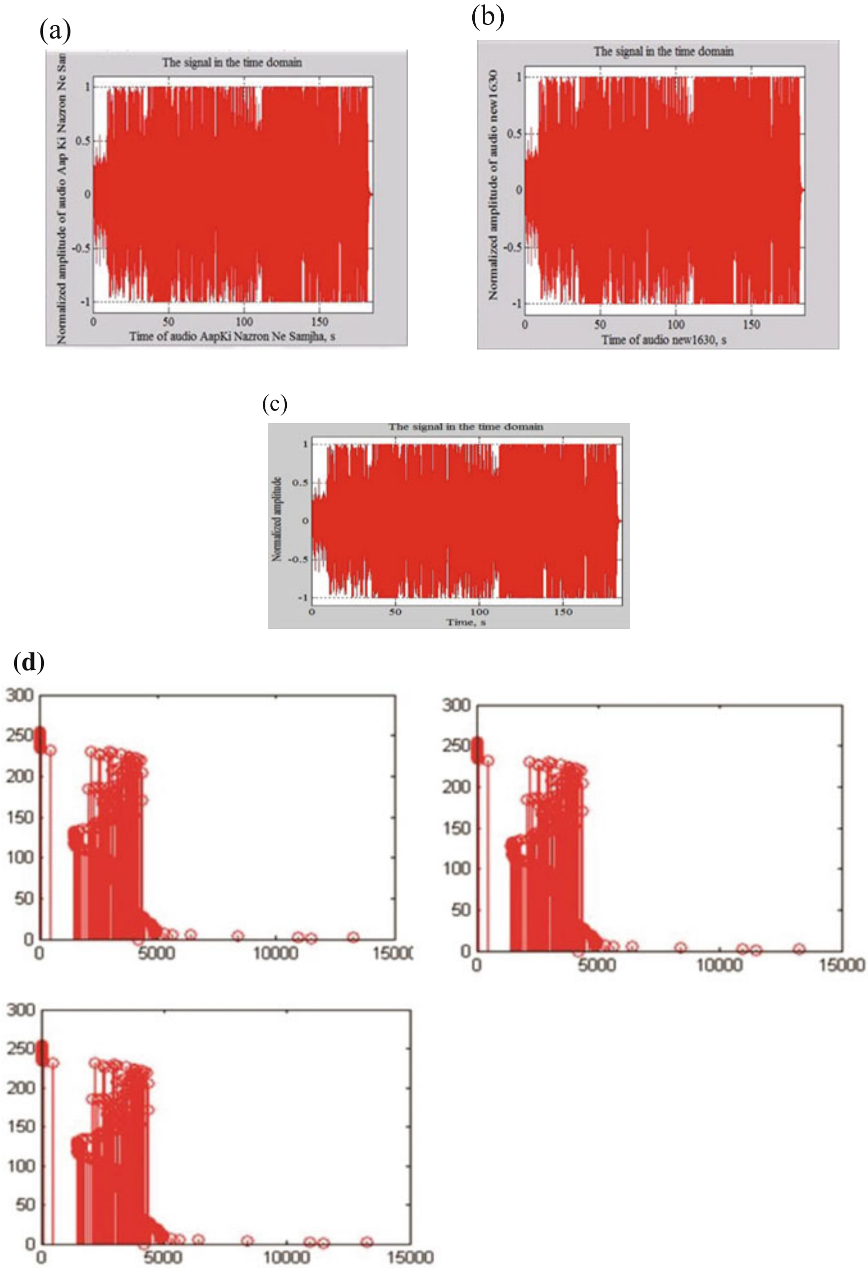


Fig. 4 **a b**. Time-domain representation of audio signal (with hidden RSA and DES ciphertext), **c** Time-domain representation of audio signal (without hidden data), **d** Histogram of image without data, with hidden RSA ciphertext and with hidden DES ciphertext

References

1. Gambhir, A., Khara, S.: Integrating RSA cryptography & audio steganography. In: IEEE, ICCCA 2016, pp. 481–483. IEEE, India (2016). https://doi.org/10.1109/ccaa.2016.7813767
2. Joseph, R., Sundaram, V.: Cryptography and steganography-a survey. IJCTA **2**(3), 626–630 (2011)
3. Krenn, R.: Steganography and steganalysis, an article, Santa Barbara, CA. (2004). http://www.krenn.nl/univ/cry/steg/article.pdf
4. Pye Pye, A., Min Tun, N.: A novel secure combination technique of steganography and cryptography. IJITMC **2**(1), 55–62 (2014)
5. Gambhir, A., Mishra, A.: CRYPTICSTEGANOGRAPHY: a new data hiding technique with multilayer security system. IJIACS **4**(1), 134–136 (2015)
6. Usha, S., Satish, K., Boopathybagan, K.: A secure triple level encryption method using cryptography and steganography. In: IEEE Computing Communication and Networking Technologies (ICCCNT), 1017–1020 (2011)
7. Gambhir, A.: RSA algorithm or DES algorithm? (a comparative analysis). JECAS **3**(4), 27–29 (2014)
8. Anuradha, H.K.: Enhanced LSB technique for audio steganography In: IEEE Computing Communication and Networking Technologies, India, pp. 89–95 (2012)
9. Arfan, S., Kirankumar, S., Vishal, U., Neeraj, V.: Audio steganography and security using cryptography. IJETAE India **4**(2), 317–319 (2014)
10. Piyush, M., Paresh, M.: Visual cryptographic steganography in images In: IEEE Computing Communication and Networking Technologies (ICCCNT), pp. 1–6 (2010)
11. Manoj, K., Amit, U., Shalini, A.: Adaptive steganographic algorithm using cryptographic encryption RSA algorithms. JEC&AS **2**(1), 1–3 (2013)
12. Deshpande, P., Sharma, S.C., Peddoju, S.K., et al.: Security and service assurance issues in Cloud environment. Int. J. Syst. Assur. Eng. Manag. (2016). https://doi.org/10.1007/s13198-016-0525-0
13. Deshpande, P., Sharma, S.C., Peddoju, S.K.: Data storage security in cloud paradigm. In: Proceedings of Fifth International Conference on Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing, vol. 436. Springer, Singapore (2016)
14. Deshpande, P., Sharma, S.C., Peddoju, S.K., Abraham, A.: Efficient multimedia data storage in cloud environment. Informatica Ljubljana **39**(4), 431–442 (2015)
15. Deshpande, P., Sharma, S.C., Kumar, P.S.: Security threats in cloud computing. In: International Conference on Computing, Communication & Automation (ICCCA), pp. 632–636 (2015)

Content-Based Image Retrieval Using Color and Texture Features Through Ant Colony Optimization

Nitin Jain and S. S. Salankar

Abstract Content-based image retrieval (CBIR) is retrieving relevant images from the large image database through visual characteristics. Each image in the database and query image is represented through feature vector derived from color and texture features in the image. These feature vectors are compared for relevance to obtain similar images in CBIR system. Therefore, length of the feature vector is very important in the CBIR system. Higher length of the feature vector increases number of comparison and in turn, increases the computational complexity, whereas lower length of the feature vector reduces comparison and complexity. In this paper, performance of the proposed CBIR system using color and texture feature extraction through histogram and Gabor wavelet transform, respectively, is presented. It is necessary to extract all the features of each image from image database and query images. These features are further presented for ant colony optimization to reduce the length of the feature vector. These final features are used in image retrieval process. Experiment results clearly show that the proposed CBIR system through ant colony optimization algorithm performance is better than other algorithms by 1.8% with respect to precision and recall. Also, the proposed algorithm clearly demonstrates the improvement by 10% on the precision and recall using only color and texture features. One of the biggest advantage and improvement was reduction in retrieval time in comparison with the other algorithms.

Keywords Content-based image retrieval · Color features · Texture features
Ant colony optimization

N. Jain (✉)
Department of Electronics Engineering, Lokmanya Tilak College of Engineering,
Navi Mumbai, India
e-mail: nitikanj@rediffmail.com; nitinjain2303@gmail.com

S. S. Salankar
Department of Electronics & Telecommunication Engineering, G. H. Raisoni
College of Engineering, Nagpur, India

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_104

1 Introduction

Sharing, searching, and posting of information, data, and multimedia data has become very easier due to tremendous growth in digital technology, especially in Internet and communication science and engineering. Multimedia data has been aggressively used in real-time audio and video conferencing, broadcast monitoring, and content-based image retrieval. The main function of the CBIR system is to retrieve similar images from the image database [1–3]. Color and texture features are largely used in CBIR system due to simplicity, higher retrieval accuracy, and reduce semantic gap [4–6]. Feature extraction is the method that extracts color and texture features from the images. Large number of methods and algorithms has been suggested in literature for CBIR systems. Recently, techniques based on face recognition [7], feature extraction and comparison [8–10], data mining [11], soft computing [12], and many more are presented for CBIR. The main objectives of the CBIR system are higher precision and recall, reduce retrieval time and semantic gap, and reducing human intervention in the retrieval process. Color and texture features extracted through color histogram and Gabor wavelet transform are found to be suitable to increase accuracy of the image retrieval system. Gabor wavelet transform effectively demonstrates the image with certain functions such as shape, textures, and edges. The primary objective of texture-based CBIR system is to retrieve images or regions with similar texture. Whenever it is desired to use texture-based images retrieval method, it is clearly understood that we are interested in images that have homogenous texture. Texture features are extracted through based on the wavelet transformation using approximation, vertical, horizontal, and diagonal frequency components of an image. The decomposed image by Haar filter consists of LL, LH, HL, and HH components. LL includes low-frequency factors and HL, LH and, HH include factors of high frequencies in horizontal, vertical, and diagonal directions, respectively. The norm of the rows of LL and LH matrices and the norm of the columns of HL matrix are computed as wavelet coefficients. Color features of the image are extracted through color histogram. It reduces complexity, computation and is invariant to rotation. These features provide measurement for color resemblance between query and database images with sufficient accuracy. These values are further applied for ant colony optimization. Then, it can be compared for similarity to the values of images indexed in a database for CBIR system. The similarity between extracted features from query and database images is considered as a measure to retrieve images. Ant colony optimized features are categorized based on feature types color and texture. Euclidean distance is used as similarity measure for texture and color statistic features. Finally, similarity measures are used for color histogram features and texture features through Gabor wavelet transform, respectively. In this paper, the performance of CBIR system based on combined color and texture features through ant colony optimization technique is presented. Visual content of the images are effectively described through Gabor filter and color histogram. Feature vector consists of concatenated texture and color features are obtained through histogram and wavelet transformation, respectively. The computational complexity due to longer feature vector length is optimized through ant colony optimization technique. It is a powerful, popular, and probabilistic method to optimized computational problems. It also assists in decreasing recognition time of the content-based image retrieval even despite the fact that image database may be too large. The paper is organized as follows: Sect. 2 describe the proposed algorithm for CBIR systems; results are discussed in Sect. 3 and finally concluded in Sect. 4.

2 Proposed Method

In this section, the proposed methodology for CBIR system using optimized color and texture features obtained through ant colony optimization is discussed. The input query image and all database images are applied for color and texture feature extraction process through histogram and Gabor wavelet transform. The derived features are then processed through ant colony optimization algorithm to obtaine optimum representation of desired features. Images in database are processed offline and the extracted optimized features are stored in feature vector for comparison and retrieval process. Ant colony optimization algorithm assists in obtaining optimum features that reduces retrieval time without scarifying on retrieval accuracy and precision.

2.1 Color Feature Extraction

Let I_q and I_d represent the input query and database image with $m \times n$ number of pixels, respectively. The steps involved in the color feature extraction through color histogram are as follows:

1. Convert RGB into HSV color space of an image.
2. Color quantization is carried out using color histogram by assigning eight levels to hue, eight to saturation, and eight to value to give a quantized HSV space with $8 \times 8 \times 8 = 512$ histogram bins.
3. Obtained normalized histogram.
4. Obtained color moments for H, S, and V color space.
5. Repeat steps 1–4 for all images in database and query image.

2.2 Texture Features Extraction

The essential steps for obtaining texture features using Gabor wavelet transform are as follows:

1. Split the input query and database image into R, G, and B color space.
2. Obtained the approximate, horizontal, vertical, and diagonal coefficients using Gabor wavelet transformation for each R, G, and B through (1).

$$G_{mn}(x, y) = \sum_{s}\sum_{t} I(x - s, y - t)g^{*}_{mn}(s, t) \quad (1)$$

where s and t are the filter mask size variables, g^{*}_{mn} is the complex conjugate of the mother Gabor function gmn, and Gmn is the convolution result corresponding to the Gabor kernel at orientation m and scale n. After applying Gabor filters on the image with different orientation at different scale, we obtain an array of magnitudes. These magnitudes represent the energy content at different scale and orientation of the image.

3. The primary objective of texture-based CBIR system is to retrieve images or regions with similar texture. Whenever it is desired to use texture-based images retrieval method, it is clearly understood that we are interested in images that have homogenous texture, therefore the following energy E_{mn}, mean μ_{mn}, standard deviation σ_{mn}, and skew α_{mn} of the magnitude of the transformed coefficients are used to represent the homogenous texture feature of the region given as follows.

$$E_{mn} = \sum_{x}\sum_{y} |G_{mn}(x, y)| \quad (2)$$

$$\mu_{mn} = \frac{E_{mn}}{P\ x\ Q} \quad (3)$$

$$\sigma_{mn} = \frac{\sqrt{\sum_{x}\sum_{y} (|G_{mn}(x, y)| - \mu_{mn})^2}}{P\ x\ Q} \quad (4)$$

$$\alpha_{mn} = \frac{(\sum_{x}\sum_{y} (|G_{mn}(x, y)| - \mu_{mn})^3)^{1/3}}{P\ x\ Q} \quad (5)$$

A feature vector F (texture representation) is created as the feature components. With M scales and N orientations used in common implementation, the feature vector is given by

$$F = (E_{mn}, \mu_{mn}, \sigma_{mn}, \alpha_{mn}, \ldots) \text{ for all m, n} \quad (6)$$

All the texture features are obtained using Eqs. 2–5 for diagonal coefficients only through M orientations and N scales. The features are arranged in single vector (F) using Eq. 6 for further retrieval of images from database.

4. Repeat steps 1–4 for all images in database and query image.

2.3 Ant Colony Optimization

1. Initialize counter x = number of images in database (x_{max}).

2. Initialized number of ants, number of features per image i_{max}, and set the intensity of pheromone trail associated with any feature.
3. Assign one ant randomly to any feature and each ant creates relation/path.
4. The feature subsets found in iterations is ordered using method providing with some weights from zero to one (initially 1).
5. Optimized feature vectors are obtained that represents image with less number of features.
6. Similarity between the query features and database features are determine through Euclidean distance method.
7. Repeat steps 3–6 entill $x < x_{max}$ else continue.
8. Precision and recall are measure to determine the accuracy of the system.

The performance evaluation of the CBIR system was evaluated through the standard procedure in terms of the two parameters precision (P) and recall (R) values. The number of relevant images in the collection is the number of images that are in the same particular category with the query image. The total number of images retrieved is the number of images that are returned by the search engine. P and R are defined as

$$P = \frac{number\ of\ relevant\ images\ retrieved}{total\ number\ of\ images\ retrieved} \tag{7}$$

$$R = \frac{number\ of\ relevant\ images\ retrieved}{total\ number\ of\ relevant\ images} \tag{8}$$

3 Experimental Results and Discussions

In this experiment, an input query image is applied to the CBIR system for retrieving relevant images from the database. Before retrieval process, all the images in the database are processed through feature extraction and ant colony optimization algorithms. The requisite feature vectors for each image are obtained and stored for further retrieval process. Similarly, every query image is applied for feature extraction and ant colony optimization algorithms to obtain feature vector. These feature vectors are then applied for comparison through Euclidean distance measurement algorithm and retrieved based on similarity index. For uniformity in all algorithms and comparison purpose, the size of all the images was fixed at 300×350 pixels. For color feature extraction process through histogram, the number of bin was set at 8. Texture feature extraction through Gabor wavelet transform was performed with 6 scales and 12 orientations. Thus, combined feature vector consist of texture features obtained from 72 components of each image and color moments through histogram in HSV space. Finally, ant colony optimization process selects eight features from the texture and four color moments for final comparison and retrieval of images. Our algorithm was implemented/executed on i5 2.4 GHz processor with 4 GB RAM through computer programming language Matlab R2015a. The image

Table 1 Precision and recall

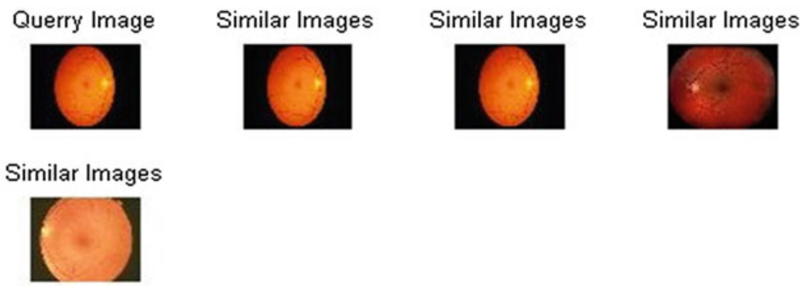
Image categories	Precision	Recall
Brain	0.72	0.93
Retina	0.69	0.93
Coins	0.87	0.95
Sun	0.825	0.94
Flowers	0.82	0.92
Average	0.785	0.934

Table 2 Comparison with other algorithms

Reference	Method	Average precision
[13]	Feature extraction without relevance feedback	0.555
[13]	Feature extraction with relevance feedback	0.747
[14]	Feature extraction with linear discriminant analysis	0.772
[15]	Color and texture fused features	0.622
[16]	Color layout and Gabor filters	0.606
[17]	Color ton distributed descriptors	0.532
[18]	Hybrid information descriptors and DCT	0.731
Our work	Feature extraction and ant colony optimization algorithm	0.785

database consists of 500 images from 5 different categories that include 100 images each of flowers/leaves, brain, retina, coins, and sun. The query image was applied from the database image and similar images were retrieved. The performance of the CBIR system was measured through two parameters precision and recall. These parameters have been used by many researchers for comparison of CBIR system. The CBIR system based on texture features and ant colony optimization algorithm was compared with other algorithms in the literature. We have previously implemented CBIR system based on texture and color feature extraction algorithm, second texture and color feature extraction algorithm with relevance feedback and third texture and color feature extraction algorithm with linear discriminant analysis. The experimental results clearly show improvement in precision and recall by 1.8% as compared to previously implemented algorithms. Also, the optimum feature vector length reduces comparison and decreases retrieval time. Table 1 shows the values of precision and recall obtained through texture and color features and ant colony optimization algorithm. Figure 1 demonstrates the relevant images retrieved through the algorithm for various categories. It is observed that the retrieval accuracy is highest when the database images are similar, whereas accuracy slightly reduces when the database consist of dissimilar images. However, the proposed CBIR system performs better as compared to other algorithms and methods. Table 2 shows the comparison of proposed algorithm with similar feature extraction algorithm with optimization.

(a) Flowers / leaves



(b) Retina

(c) Brain

Fig. 1 Retrieved images for CBIR system

4 Conclusion

In this paper, we proposed CBIR system based on combined color and texture feature extraction algorithm through ant colony optimization. An experimental result obtained through simulation on various images clearly shows improvement in average precision as compared to other algorithms. Ant colony optimization algorithm was found suitable for CBIR system that effectively reduces feature vector length to an optimum value. Reduction in feature vector length does not affect the accuracy of the CBIR system. Thus, it clearly benefits in comparison process without affecting accuracy. It also enhances the prospect of designing system with combining various features such as shapes, texture, and color. Also, optimization process may be helpful in region-based CBIR system due to smaller length of feature vectors. Further, the algorithm can be combined with relevance feedback mechanism to enhance accuracy and reduce semantic gap.

References

1. Mahamuni, C.V., Wagh, N.B.: Study of CBIR methods for retrieval of digital images based on colour and texture extraction. In: IEEE International Conference on Computer Communication and Informatics, Coimbatore, India (2017)
2. Gerard, S., Buckely, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1998)
3. Chen, Y., Wang, J.: Image categorization by learning and reasoning with regions. J. Mach. Learn. Res. **5**, 913–939 (2004)
4. Long, F., Zhang, H., Dagan, H., Feng, D.: Fundamentals of content based image retrieval. In: Multimedia Information Retrieval and Management. Technological Fundamentals and Applications, Multimedia Signal Processing Book, Chapter 1. Springer, Berlin, Heidelberg New York, pp. 1–26 (2003)
5. Manjunath, B., Ma, W.: Texture features for Browsing and retrieval of image data. IEEE Trans. Pattern Anal. Mach. Intell. **18**(8), 837–842 (1996)
6. Gevers, T., Smeulders, A.: Pictoseek: combining color and shape invariant features for image retrieval. IEEE Trans. Image Process. **9**(1), 102–119 (2001)
7. Muhammad Fachrurrozi, E., Saparudin, M.: Multi-object face recognition using content based image retrieval (CBIR). In: IEEE International Conference on Electrical Engineering and Computer Science, Indonesia (2017)
8. Fuertes, J., Lucena, M., Perez, N., Martinez, J.: A scheme of color image retrieval from databases. Pattern Recogn. Lett. **22**, 323–337 (2001)
9. Ouyang, A., Tan, Y.: A novel multi-scale spatial-color descriptor for content based image retrieval. In: Proceedings of the 7th International Conference on Control, Automation, Robotics and Vision, Mexico, vol. 3, pp. 1204–1209 (2002)
10. Yu, H., Li, M., Zhang, H., Feng, J.: Color texture moments for content-based image retrieval. In: Proceedings of the International Conference on Image Processing, Rochester, New York, USA, Sept 22–25, vol. 3, pp. 929–932 (2002)
11. Guan, H., Wada, S.: Flexible color texture retrieval method using multiresolution mosaic for image classification. In: Proceedings of the 6th International Conference on Signal Processing, vol. 1, Feb 2002, pp. 612–615
12. Lew, M., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: state of the art and challenges. ACM Transactions on Multimedia Computing, Communications and Applications **2**(1), 1–19 (2006)
13. Jain, N., Salankar, S.S.: Performance estimation of relevance feedback for content based image retrieval system. In: IEEE International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing, Vellore, India (2018)
14. Jain, N., Salankar, S.S.: Content based image retrieval using improved gabor wavelet transform and linear discriminant analysis. In: IEEE International Conference on Convergence in Technology, Pune, India (2018)
15. Yue, J., Li, Z., Liu, L., Fu, Z.: Content-based image retrieval using colour and texture fused features. In: Mathematical Computational Modeling, vol. 54, pp. 1121–1127 (2011)

16. Jalab, H.A.: Image retrieval system based on colour layout descriptor and Gabor filters. In: IEEE Conference on Open System (ICOS) Langkawi, Malaysia (2011)
17. Rahimi, M., Moghaddam, M.E.: A content based image retrieval system based on colour ton distributed descriptors. In: Signal Image and Video Processing, (SIViP), vol. 9, pp. 691–704 (2013)
18. Singh, V.P., Malhotra, S., Srivastava, R.: Combining hybrid information descriptors and DCT for improved CBIR performance. In: International Conference on Control, Computing, Communication and Materials (ICCCCM), Allahabad, pp. 1–5 (2016)

Correction to: Performance of Internal Cluster Validations Measures For Evolutionary Clustering

Pranav Nerurkar, Aruna Pavate, Mansi Shah and Samuel Jacob

Correction to: Chapter "Performance of Internal Cluster Validations Measures For Evolutionary Clustering" in: B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*, Advances in Intelligent Systems and Computing 810, https://doi.org/10.1007/978-981-13-1513-8_32

In the original version of the book, the following belated corrections have been incorporated in chapter "Performance of Internal Cluster Validations Measures For Evolutionary Clustering": The affiliation "Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu, Rajasthan, India" of author "Mansi Shah" has been changed to "Department of CE & IT, Rizvi CoE, Mumbai, India". Similarly, the affiliation "Department of CE & IT, Rizvi CoE, Mumbai, India" of author "Samuel Jacob" has been changed to "Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu, Rajasthan, India". The city "Jhunjhunu" in the affiliation "Department of CE & IT, Atharva CoE, Jhunjhunu, India" of author "Aruna Pavate" has been changed to "Mumbai". The correction chapter and the book have been now updated with the changes.

The updated original online version for this chapter can be found at
https://doi.org/10.1007/978-981-13-1513-8_32

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_105

Author Index

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8